Evaluation of Climate Models For Seasonal Forecasting in the MENA Region

Prepared by: **Berrahmouch Nohayla and Mohamed El-Badri**
Hassania School of Public Works , Casablanca, Morocco

Supervised by: **Mrs. Wafae Badi and Mr. Nicholas Savage**
Direction Générale de la Météorologie, Morocco (Wafae Badi)
Met Office, Exeter, UK (Nicholas Savage)

2024 - 2025

Contents

ACKNOWLEDGMENTS

We would like to express our deepest gratitude to **Mrs. Wafae Badi** for her unwavering support and invaluable guidance throughout this project. Her insightful counsel and constant encouragement enabled us to overcome various challenges and maintain our focus. Her dedication to fostering progress, coupled with her constructive feedback and open-minded approach, were instrumental in shaping the direction of this work. Her mentorship has truly been a cornerstone of our journey, and we are profoundly grateful for her invaluable contributions.

Special thanks go to **Mr. Nicholas Savage** and his exceptional team at the UK Met Office. Their generosity in sharing their expertise and resources provided us with unparalleled opportunities to broaden our understanding of climate modeling. The engaging discussions and valuable insights shared by Mr. Savage and his team not only enriched this project but also fueled our motivation to explore innovative avenues. Their commitment to advancing climate science inspired us to aim higher and achieve more.

We are also immensely grateful to **Mr. Bari**, whose dedicated supervision, thoughtful suggestions, and constructive critiques significantly enhanced the quality of this work. His ability to balance critical feedback with motivating encouragement made a remarkable difference, guiding us through challenging moments and ensuring steady progress.

In addition, we would like to acknowledge the support received through the **WISER MENA project**. **Nicholas Savage's time was funded via the WISER MENA project.** The Weather and Climate Information Services (WISER) Programme is funded with UK International Development from the UK government and led by the Met Office in the UK. This work has been partially supported by UK International Development from the UK government; however, the views expressed do not necessarily reflect the UK government's official policies.

Lastly, we extend our heartfelt appreciation to all those who, directly or indirectly, contributed to this project. Your cooperation, guidance, and belief in our work have made this journey a fulfilling and enlightening experience. While this project is a testament to hard work and collaboration, it is also a reflection of the collective effort and support of everyone who believed in its success. To you, we owe our sincere thanks.

PREFACE

The MENA seasonal forecasting models have undergone both probabilistic and deterministic evaluations. This research study is regarded as the pioneering work and the first of its kind in this area which helps in situational context improvement in seasonal forecasting models. Given the alarming rate of increase in the impacts caused by extreme climatic events including severe droughts, and extreme heat and other climate sensitive issues in the MENA region, this work is a key contribution towards alleviating these issues=

Due to climatic extremes in the MENA region, agriculture, human livelihood, and natural resources are heavily affected. Consequently, it has become almost necessary to have forecasts of seasons that are credible so as to characterize the impacts, or to enhance preparedness. Although seasonal forecasting models have been widely researched and practiced in many parts of the world, their use in MENA countries' local level remains scarce. This gap is resolved in this study, providing new knowledge and tools for climate scientists working in the region.

In this work, we intend to broaden the knowledge fabric of climate change science by focusing on the climate change and variability vulnerability of the MENA region. The results obtained not only improve the comprehension of the dynamics of the local climate, but also lays a framework for specific approach to be employed for adaptation strategies.

We are immensely grateful to every individual or organization who has helped support this project and guided us through uncharted territory in the spectrum of MENA climate predictions.

OVERVIEW AND RATIONALE OF THE STUDY

The last couple of decades have witnessed a surge in demand for seasonal climate forecasting. Global advancements in space science and technology have lead to the better anticipation of climate seasons up to a through range of 3-12 months. This is crucial for effective planning in major industries like agriculture or energy management, among others. These advancements breed an increased dependence on seasonal forecasting and in turn create a higher demand for accurate forecasting mechanisms. Therefore two central methodologies have witnessed prominence – deterministic and probabilistic methods. A hindsight understanding of these mechanisms is imperative, as they are useful for evaluating and understanding the shortcomings and effectiveness of different models employed in forecasting seasonal amps.

Probabilistic forecasts take one step forward, do not try to predict an ideal scenario and present different potential outcomes, each with a defined probability. Efforts, though different, instruct towards the same ends; meeting a specific operational/strategic need. Lorenz's butterfly effect presents the case for one such endeavor- it shows how a non-linear system's response can drastically alter depending on the initial conditions. Such chaos is especially present in weather and climate systems where even the slightest details can have large ramifications over longer periods.

The study on the other hand tries to develop such relationships that integrate conceptual developments in seasonal forecasting efforts with applicable methods.

CHAPTER 1

INTRODUCTION

1.1 Context

1.1.1 Overview of Climate Modeling and Seasonal Forecasting

Climate modeling is the process of using mathematical representations of the Earth's atmosphere, oceans, land surface, and ice systems to simulate and predict climate dynamics. These models are based on fundamental physical principles, such as the conservation of mass, energy, and momentum, and are implemented through numerical methods that solve complex equations governing the interactions between these systems.[1] Climate models range from global circulation models (GCMs), which simulate large-scale atmospheric and oceanic processes, to regional climate models (RCMs), which provide localized projections by incorporating finer-scale topographic and land-use details.[2] Seasonal forecasting, a subset of climate modeling, refers to the prediction of climate conditions, such as temperature and precipitation, over a period of one to six months. These forecasts rely on initial conditions (e.g., sea surface temperatures, soil moisture) and slowly varying components of the climate system, such as oceanic or atmospheric anomalies like the El Niño-Southern Oscillation (ENSO).[3] The basic principle behind seasonal forecasting is to leverage these slowly varying components, which have a predictable influence on regional weather patterns, using ensemble simulations to quantify uncertainties and provide probabilistic predictions.[4]

Seasonal forecasts play a crucial role in decision-making and planning across various sectors, including agriculture, water management, and climate risk mitigation. These forecasts provide early warnings of high-impact climate scenarios, enabling proactive decisions that result in financial savings, risk reduction, and optimized resource use. For instance, in agriculture, they assist farmers in selecting appropriate crops and determining optimal planting times based on anticipated water availability, thereby mitigating risks associated with droughts or excessive rainfall.[5]

[1] McGuffie, K. and Henderson-Sellers, A., 2014. A Climate Modelling Primer. https://doi.org/10.1002/9781118687853

[2] Flato et al., 2013. Evaluation of Climate Models. IPCC AR5 Chapter 9. https://www.ipcc.ch/report/ar5/wg1/chapter-9-evaluation-of-climate-models/

[3] Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R., 2013. Seasonal climate predictability and forecasting: Status and prospects. https://doi.org/10.1038/ngeo1714

[4] Palmer, T. N., & Anderson, D. L., 1994. The prospects for seasonal forecasting—a review paper. https://doi.org/10.1256/smsqj.50402

[5] Werner, M. and Linés, C., 2024. Seasonal forecasts to support cropping decisions. https://doi.org/10.5194/egusphere-egu24-13436

Seasonal forecasts also support pre-harvest strategies, such as hedging decisions, which help shield farmers from price volatility, although their adoption is often hindered by perceptions of inaccuracy and complexity.[6] In water management, seasonal forecasts are vital for mitigating drought impacts, particularly in semi-arid regions, by enabling improved reservoir operations and efficient water allocation to reduce losses.[7] Additionally, these forecasts, when linked to hydrological models, improve predictions of water balance and inform critical decisions regarding water storage and distribution, despite occasional discrepancies between predicted and desired variables.[8] Seasonal forecasts are increasingly applied in climate risk management, where they help predict extreme weather events, providing decision-makers with tools to minimize societal and economic damages.[9] For example, accurate predictions of heatwaves or floods allow authorities to implement adaptive measures, reducing infrastructure damage and safeguarding public health. In economic sectors such as energy and water management, tailored seasonal forecasts enhance decision-making efficiency by aligning forecasts with user needs, thereby optimizing outcomes.[10] Despite their significant potential, the effectiveness of seasonal forecasts depends on their accuracy, relevance to user needs, and ease of use. Improved communication, stakeholder training, and efforts to bridge the gap between forecast complexity and user understanding are essential to maximize their utility.

1.1.2 Importance of Seasonal Climate Forecasts in MENA

Seasonal climate forecasts are critically important across the MENA region, where high temperatures, low water availability, and vulnerability to climate variability create substantial challenges for sustainable development. Forecasts provide early warnings of droughts, heatwaves, and other extreme weather events, enabling decision-makers to implement proactive measures to mitigate impacts on water resources, agriculture, and infrastructure.[11] In agriculture, these forecasts help farmers optimize crop selection and planting schedules, reducing the risks of crop failure in this water-scarce region.[12] In the water sector, seasonal forecasts guide reservoir management by predicting rainfall variability, improving water storage strategies, and ensuring more equitable water distribution.[13] With increasing climate risks, these forecasts also support disaster risk management by allowing governments to prepare for extreme events, such as heatwaves and floods, which are becoming more frequent in the region due to climate change.[14] Moreover, the economic benefits of using seasonal forecasts are significant. By enabling energy companies to anticipate peak demand periods driven by heatwaves, and by helping municipalities optimize water usage during droughts, these forecasts provide cost savings and efficiency gains.[15] However, challenges persist in ensuring the accuracy and usability of these forecasts. The arid and semi-arid nature of much of the MENA

[6] Hunt et al., 2020. Seasonal Forecast Based Preharvest Hedging. https://doi.org/10.22004/AG.ECON.309761

[7] Portele et al., 2021. Seasonal forecasts offer economic benefits for hydrological decision-making. https://doi.org/10.1038/s41598-021-89564-y

[8] MacLeod et al., 2023. Translating seasonal climate forecasts into water balance forecasts. https://doi.org/10.1371/journal.pclm.0000138

[9] Castino et al., 2023. Towards seasonal prediction of extreme temperature indices. https://doi.org/10.5194/ems2023-590

[10] Goodess et al., 2022. The Value-Add of Tailored Seasonal Forecast Information. https://doi.org/10.3390/cli10100152

[11] Dunn et al., 2020. The changing climate of MENA. https://pubs.giss.nasa.gov/abs/gu00200u.html

[12] Werner, M., and Linés, C., 2024. Seasonal forecasts to support cropping decisions. https://doi.org/10.5194/egusphere-egu24-13436

[13] Portele et al., 2021. Seasonal forecasts for hydrological decision-making. https://doi.org/10.1038/s41598-021-89564-y

[14] Castino et al., 2023. Towards seasonal prediction of extreme temperature indices. https://doi.org/10.5194/ems2023-590

[15] Goodess et al., 2022. Value-Add of tailored seasonal forecast information. https://doi.org/10.3390/cli10100152

region, coupled with complex interactions between regional climate drivers, makes it difficult to provide highly localized forecasts.[16] Addressing these challenges through improved modeling techniques and stakeholder engagement will be critical to maximizing the value of seasonal forecasts in the MENA region, ensuring better preparedness and resilience against a changing climate.

1.2 Objectives of the Work

The primary objective of this work is to evaluate the effectiveness of climate models, focusing specifically on their performance in predicting key climate variables such as temperature, precipitation. This evaluation incorporates both deterministic and probabilistic approaches to identify the most skillful models and their suitability for practical applications.

1.2.1 Specific aims of evaluating deterministic and probabilistic models.

The evaluation of deterministic and probabilistic models is essential for understanding their unique strengths, limitations, and potential applications in diverse fields. Deterministic models, which generate a single, precise outcome based on initial conditions, are widely used when exactness and reproducibility are critical, such as in engineering and physical simulations.[17] Their evaluation focuses on assessing accuracy and reliability under specific conditions, providing clarity in cause-and-effect relationships. In contrast, probabilistic models incorporate uncertainty by assigning probabilities to various potential outcomes, enabling the representation of real-world complexities and variability.[18] These models are particularly beneficial for strategic planning and risk management, where understanding a range of possible scenarios is crucial. The evaluation of both types of models includes conducting sensitivity analyses to determine how changes in input variables affect outcomes, which helps in identifying key drivers of uncertainty and improving model performance.[19] Additionally, risk assessment is a vital component, with deterministic approaches offering straightforward estimations for defined scenarios, while probabilistic approaches address uncertainties by simulating a spectrum of possible outcomes.[20] These evaluations also aim to support decision-making processes by identifying which type of model is more appropriate for specific contexts—deterministic models for precise predictions and probabilistic models for flexible planning under uncertainty.[21] Finally, probabilistic models are often recognized for their adaptability in dynamic environments, as they can incorporate new data and adjust probability distributions to reflect evolving conditions, making them indispensable for complex systems where deterministic models may fall short.[22] Together, the evaluation of deterministic and probabilistic models provides invaluable insights into their suitability for addressing specific challenges, supporting informed decision-making, and advancing model development.

[16]Latif et al., 2011. ENSO predictability and regional climate impacts. https://doi.org/10.1175/2010JCLI3405.1

[17]McGuffie, K., and Henderson-Sellers, A., 2014. *A Climate Modelling Primer*. Wiley. https://doi.org/10.1002/9781118687870

[18]Palmer, T., and Hagedorn, R., 2006. *Predictability of Weather and Climate*. Cambridge University Press. https://doi.org/10.1017/CBO9780511617652

[19]Seneviratne, S.I., et al., 2021. *Metrics for climate model evaluation: A review*. Nature Communications. https://doi.org/10.1038/s43247-021-00094-x

[20]PreventionWeb, 2021. *Deterministic and Probabilistic Risk*. https://www.preventionweb.net/understanding-disaster-risk/key-concepts/deterministic-probabilistic-risk

[21]Goodess, C.M., et al., 2022. *The Value-Add of Tailored Seasonal Forecast Information for Industry Decision Making*. Climate. https://doi.org/10.3390/cli10100152

[22]Latif, M., and Keenlyside, N., 2011. *El Niño/Southern Oscillation Predictability*. Journal of Climate. https://doi.org/10.1175/2010JCLI3405.1

1.2.2 Description of Content

This report is designed to provide a comprehensive analysis of climate model evaluation, focusing on both deterministic and probabilistic approaches. The structure of the report follows a logical progression, starting with an introduction to the fundamental concepts behind climate models. The first chapter lays the groundwork for understanding the key differences between deterministic and probabilistic models, describing how each approach is used to simulate climate systems and predict future outcomes. The methodology chapter follows, detailing the specific techniques employed to assess the models. This includes the use of both deterministic and probabilistic metrics such as Root Mean Square Error (RMSE), Anomaly Correlation Coefficient (ACC), and Brier Score, which are critical for evaluating the models' accuracy and performance in predicting climate variables like temperature and precipitation.

Next, the report moves on to the results and analysis, where the performance of the selected models is presented and compared. This chapter highlights the models' strengths and weaknesses, providing insight into how well they predict climate patterns across various geographical regions and time periods. Special attention is given to the models' skill in forecasting extreme weather events, which are particularly relevant to sectors like agriculture, water resource management, and disaster risk reduction.

The final chapter of the report provides conclusions and recommendations based on the analysis. This chapter synthesizes the findings, offering practical suggestions for improving the accuracy, usability, and application of climate forecasts. Recommendations also address how future developments in climate modeling can better meet the needs of decision-makers and stakeholders. The report as a whole seeks to contribute valuable insights into the ongoing development of climate prediction systems, aiming to enhance their effectiveness in real-world applications.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of Climate Models

2.1.1 Deterministic Models

Deterministic models rely on mathematical equations that describe the physical processes of the atmosphere. These models use fixed initial conditions to provide precise predictions, making them suitable for short-term forecasting. However, due to the chaotic nature of atmospheric systems, as demonstrated by Lorenz's theorem, deterministic models are limited in their ability to predict long-term outcomes. Small errors in initial conditions can lead to significant differences in results, reducing their reliability for seasonal or long-term forecasting.[1]

Deterministic climate models operate based on fixed initial conditions and mathematical equations that simulate physical processes in the atmosphere. These models are particularly useful for short-term predictions as they provide precise and singular forecasts. However, deterministic models are significantly limited when forecasting over extended periods. This limitation arises due to the inherent sensitivity of atmospheric systems to initial conditions—a concept known as the *butterfly effect*, introduced by Edward Lorenz in 1963. His research demonstrated that even minute changes in the initial conditions of a system could lead to vastly different outcomes over time, emphasizing the chaotic nature of weather systems.

For seasonal forecasting, deterministic models often fail because minor errors in the initial conditions can amplify, resulting in inaccurate predictions for longer timescales. Despite these challenges, deterministic models are vital for understanding specific phenomena over shorter durations with high spatial and temporal resolution.

2.1.2 Probabilistic Models

Probabilistic models address the limitations of deterministic approaches by incorporating uncertainty into forecasts. Instead of producing a single outcome, these models generate a range of possible scenarios, each with an associated probability, using ensemble simulations or statistical techniques. This makes probabilistic models particularly useful for medium- to long-term forecasts and risk assessment in climate-sensitive sectors such as agriculture, water management, and disaster

[1]Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences, 20*(2), 130–141.

mitigation.[2]

The evaluation of probabilistic models relies on metrics that assess their ability to represent uncertainty and provide actionable insights:

- **Reliability:** Measures how well predicted probabilities align with observed frequencies.
- **Resolution:** Assesses the model's ability to distinguish between different outcomes.
- **Discrimination:** Evaluates the model's ability to separate events from non-events.[3]

Probabilistic models are especially valuable for decision-making under uncertainty, as they provide stakeholders with a clearer understanding of risks and potential scenarios, enabling proactive measures to mitigate impacts.

Comparison of Deterministic and Probabilistic Models

Deterministic and probabilistic models serve complementary roles in climate modeling and forecasting. Their distinct features and applications are summarized in Table **??**.

Table 2.1: Comparison of Deterministic and Probabilistic Models

Feature	**Deterministic Models**	**Probabilistic Models**
Predictability	Produces a single fixed outcome based on initial conditions	Generates a range of outcomes with associated probabilities
Sensitivity to Initial Conditions	Highly sensitive, leading to reduced accuracy over long timeframes	Less sensitive due to ensemble techniques reducing error amplification
Application Domain	Suitable for short-term, high-resolution tasks, e.g., extreme event analysis	Ideal for medium- and long-term decision-making under uncertainty
Use of Historical Data	Limited emphasis on historical variability	Extensively relies on historical data for statistical projections
Examples	Global Circulation Models (GCMs), Regional Climate Models (RCMs)	Ensemble forecasting, statistical downscaling

While deterministic models are preferred for precise and short-term predictions, probabilistic models provide critical insights into the likelihood of various scenarios, making them indispensable for managing climate-related risks.

[2]World Meteorological Organization (2024). *Guidance on Verification of Operational Seasonal Climate Forecasts*. https://library.wmo.int/records/item/56227-guidance-on-verification-of-operational-seasonal-climate-forecasts

[3]Rapport de projet 2024–2025, *3rd Year Meteorology Modeling Project*.

2.2 STUDIES IN "MENA".

2.2.1 The current and changing climate in MENA

Much [4] of the MENA region is characterised by high temperature and low water availability, a combination of variables that have the potential to lead towards the environmental limits/threshold for safe human habitation. This makes the region particularly vulnerable to climate change and climate variability, as small variations in climate can easily produce high temperatures or extensive droughts that are harmful to human lives and livelihoods.

Changes in temperature and rainfall patterns have already been observed in the region and are expected to change further in the near future, especially if global warming exceeds 1.5 to 2 °C above the pre-industrial level. Annual mean temperatures across the MENA region have increased between 0.3–0.5°C per decade1 over the period 1980–2015 [5]. Since the 1950s, hot and cold extremes have become warmer, the number of cold days has decreased, and the number of warm days has increased (Dunn et al., 2020). There has been an increase in heat waves intensity, frequency and duration [6]. Annual mean precipitation shows a high level of spatial variability over the MENA region. During the period 1980–2015 there have been downward trends in mean annual precipitation [7] . Dry conditions, drought intensity and frequency has increased in the past over the region [8]

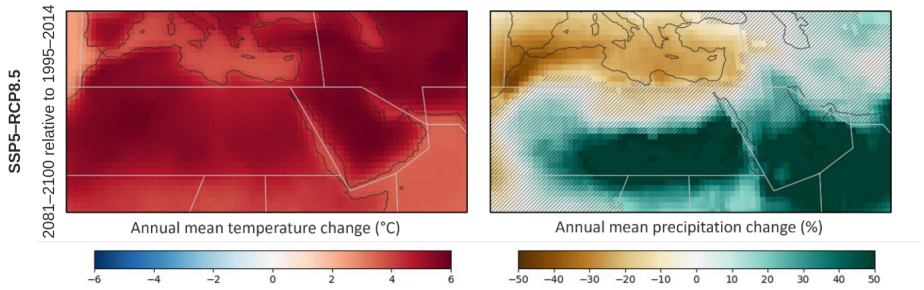

Figure 2.1: Projected change in annual mean temperature (left) and annual mean precipitation (right) between 1995–2014 and 2081–2100 under the SSP5–RCP8.5 scenario based on CMIP6 models (Gutiérrez et al., 2021). Note that precipitation change is given as a percentage: the large increases projected over Sahara and Arabian deserts equate to only a few millimetres of additional rainfall.

2.2.2 Impact-Based Evaluation

Impact-based forecasting refers to a type of weather or climate forecasting that goes beyond predicting the meteorological parameters (e.g., temperature, rainfall, wind speed) and instead focuses on predicting the potential impacts of those conditions on society, infrastructure, and ecosystems. The goal is to provide actionable insights that help communities and decision-makers prepare for and mitigate the effects of extreme weather and climate events.

[4] Met Office WISER Report
[5] (Gutiérrez et al., 2021)
[6] (Perkins-Kirkpatrick and Lewis, 2020)
[7] (Gutiérrez et al., 2021)
[8] (Seneviratne et al., 2021).

Evaluation of Seasonal Forecast Models
An impact-based evaluation[9] [10] was conducted as global study on five seasonal forecast models to identify the most effective for extreme precipitation forecasting (focuses on regions which were vulnerable to wildfire and flooding). The models assessed included:

- Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC: version 35),
- Deutscher Wetterdienst (DWD: version 21),
- Environment and Climate Change Canada (ECCC: version 3),
- Météo-France (version 8),
- UK Met Office (UK-Met: version 601).

The findings highlighted the ***UK-Met*** and ***Météo-France*** models as consistently superior across all four seasons. Meanwhile, the ECCC and CMCC models exhibited strong performance on specific indices and in particular regions, ranking just below the top two models.

ROC Scores and Regional Performance
The ROC scores indicate that forecast models perform exceptionally well in tropical and subtropical regions. This result is consistent with our study and can be attributed to the general predictability of oceanic conditions and the influence of climate drivers such as the El Niño-Southern Oscillation (ENSO). The Météo-France and UK-Met models exhibited superior performance during the SON and MAM seasons.

However, the prevalence of grids with no discrimination ROC categories is more common in extratropical regions. This can be attributed to:

- Lower predictability of extratropical variations,
- Model limitations in capturing interactions between tropical and extratropical regions,
- Challenges in representing land surface processes (De Andrade et al., 2019).

The CMCC, DWD, and ECCC models often fail to detect extreme events in many extratropical areas, underscoring the stronger performance of the UK-Met and Météo-France models in these scenarios.

Percent Bias Analysis
The analysis of Percent Bias across four seasons demonstrates a consistent underestimation by forecast models for most extreme wet precipitation indices. Key observations include:

- Forecast models underestimate extreme wet precipitation indices while overestimating light precipitation.
- Models perform better in capturing the intensity and magnitude of extreme events (e.g., highest daily and multi-day rainfall) compared to the frequency of wet or dry days.

In tropical and subtropical regions, models like ***UK-Met*** and ***Météo-France*** exhibit strong performance due to their ability to capture large-scale climate patterns. In contrast, extratropical regions show higher biases, reflecting challenges in modeling complex interactions and seasonal variations.

[9] https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2024EF004936
[10] Zahir Nikraftar, Rendani Mbuvha, Mojtaba Sadegh, Willem A. Landman

Global Model Comparison

The ***UK-Met*** model consistently demonstrates lower biases and stronger performance globally compared to the ***Météo-France*** model, highlighting its effectiveness in representing climate patterns. However, all models show limitations in accurately modeling persistent extreme wet and dry periods, particularly in extratropical areas.

2.2.3 SYSTEM 7 FRANCE

seasonal forecasting evaluation has been the subject of numerous studies, with a focus on improving the accuracy and reliability of predictions related to precipitation and other weather parameters. One such study[11] conducted a probabilistic evaluation of seasonal precipitation re-forecasting from May to November over a period of 23 years (1993–2015). The study utilized the Brier Score (BS) and its decomposition to assess forecast performance, with the aim of providing more reliable and actionable predictions for extreme weather events.

The evaluation was conducted on the operational seasonal forecasting system of Meteo-France, which used 25 ensemble members, perturbed model dynamics, and initial conditions. The system aimed to provide a more detailed probabilistic forecast, in addition to existing deterministic metrics, for both seasonal and intra-seasonal forecasts. The BS was estimated using tercile probabilities and a non-parametric counting estimator, with the GPCP[12] observation data serving as the reference.

Multiple analyses were performed to evaluate the robustness of the BS score, revealing that spatial distributions of the BS can vary significantly based on the sampling methods, reference data, and ensemble types used. The analysis showed that large errors, especially in the tropical ocean, could be reduced by using hindcast ensemble climatological samples. In particular, errors over the Nino region in the Pacific Ocean could be mitigated using these methods. This highlights the importance of employing various ensemble data sources and reference climatology to enhance the reliability of seasonal forecasts.

A notable finding was the reduction in BS when using ensemble observations, especially in the tropical ocean, suggesting that increasing ensemble size can improve forecast accuracy up to a point. However, this was not the case in all regions, as some areas, such as the tropical Indian Ocean, exhibited high BS even with different analysis methods. The study also found that intra-seasonal analyses showed similar patterns to seasonal hindcasts, but with higher BS due to reduced sample sizes, highlighting the need for higher-resolution models and improved initial conditions.

The study concluded that, despite improvements, probabilistic forecasting still faces challenges, particularly in the tropical regions, where errors fluctuate with lead time. The study emphasized the need for continued development of forecasting methods, particularly in reducing uncertainties in evaluation scores. Future evaluations should expand beyond the BS to include other metrics, such as the forecast skill score and the relative operating characteristic (ROC), to better assess forecast performance and identify system deficiencies.

This study's findings underline the importance of ensemble forecasting and the use of diverse data sources to improve the accuracy of seasonal precipitation forecasts, particularly in tropical regions where predictability remains challenging.

[11] https://www.mdpi.com/2674-0494/1/3/16

[12] Global Precipitation Climatology Project (GPCP)

CHAPTER 3

METHODOLOGY

3.1 DATA

The hindcast data used in this study was obtained using the OSOP package[1], a tool developed by the UK Met Office to facilitate the retrieval of climate and meteorological data. The dataset comprises monthly mean seasonal forecasts for temperature over the MENA (Middle East and North Africa) region.

The hindcast data spans the common period 1993–2016 and was downloaded from the Copernicus Climate Change Service (C3S) platform.

The data was retrieved for the following configurations:

- Variables: 2-meter air temperature (t2m) and total precipitation (tp).
- Forecast Range: Lead times of interest (1–3 months), it includes DJF[2], JJA[3], MAM[4], SON[5]
- Geographical Area: MENA region.
- Temporal Coverage: 1993–2016
- the used centers are $UKMO_602, ECMWF_51, ECCC_2, ECCC_3, CMCC_35, Meteo-France_8, DWD_21$

In addition to the hindcast data, this study utilized ERA5 reanalysis data, a state-of-the-art atmospheric reanalysis product produced by the European Centre for Medium-Range Weather Forecasts (ECMWF).

3.2 Deterministic Metrics

3.2.1 Spearman rank correlation

Spearman's correlation is a non-parametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function (whether linear or not).

[1] https://github.com/OSFTools/osop/tree/main/scripts
[2] December,January,February
[3] June,July,August
[4] March,April,May
[5] September,October,November

$$r_s = \frac{cov(R[H], R[O])}{\sigma_{R[H]} \cdot \sigma_{R[O]}}$$

where :

- r_s : spearman rand correlation
- H : the Hindcast.
- O : the Observation.
- R[x] : the rank of the variable x.
- σ_x : standard deviation of the variable x.

3.2.2 RMSE

$RMSE$ measures the average difference between a the hindcast and the observation.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(H_i - O_i)^2}$$

where :

- H : the Hindcast.
- O : the observation.
- i : the valid time.

3.2.3 Coefficient of Determination (R^2)

The coefficient of determination, R^2, is a statistical measure used to evaluate the goodness of fit of a model. It indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s). A value of R^2 close to 1 suggests that the model explains a large portion of the variance, while a value close to 0 indicates a weak relationship.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - H_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}$$

where:

- R^2: Coefficient of determination.
- H_i: Predicted value (Hindcast).
- O_i: Observed value (Observation).
- $\bar{O}$: Mean of the observed values.
- $\sum_{i=1}^{n}(O_i - H_i)^2$: Residual sum of squares (unexplained variance).
- $\sum_{i=1}^{n}(O_i - \bar{O})^2$: Total sum of squares (total variance).

3.3 Probabilistic Metrics

In the WMO[6] . Guide, several criteria are provided for evaluating a good forecast. Each criterion offers insight into specific aspects of the model but cannot, on its own, fully determine the forecast's quality. By combining all the criteria, we can comprehensively assess the performance of the model.

3.3.1 Resolution

Resolution measures whether the outcome differs given different forecasts, while discrimination measures whether the forecasts differ given different outcomes.

Discrimination looks at how well your forecast separates cases when the event (outcome) happens (pass) from when it doesn't happen (fail). It's about telling apart the events. Resolution looks at how well your forecast adapts to different situations, giving distinct probabilities for different cases. It's about adjusting to the situation.

Resolution measures how well a forecast distinguishes between different outcomes. A forecast has high resolution if the predicted probabilities vary significantly depending on the actual outcome. In other words, resolution tells us whether the forecast changes (e.g., gives different probabilities) when the actual outcome changes. High resolution: The forecast gives distinct and varying probabilities when different events (outcomes) occur. For example, if in one case the forecast predicts a high probability of rain and it rains, and in another case predicts a low probability and it doesn't rain, the forecast shows good resolution. Low resolution: If the forecast probabilities don't change much regardless of whether it rains or not (e.g., always predicting a 50% chance of rain), the forecast has poor resolution because it fails to capture the differences in actual outcomes. Resolution can be determined by measuring how strongly the outcome is conditioned upon the forecast. If the outcome is independent of the forecast, the forecast has no resolution and is useless Forecasts with no resolution are neither "good" nor "bad", but are useless. Metrics of resolution distinguish between potentially useful and useless forecasts, but not all these metrics distinguish between "good" and "bad" forecasts.

The following equation represents the "resolution" component of the Brier Score (BS) decomposition, which quantifies how well a set of probability forecasts differentiates between events and non-events:

$$\textbf{Resolution} = \frac{1}{n} \sum_{k=1}^{d} n_k \left(\bar{y}_k - \bar{y}\right)^2 \tag{3.1}$$

where:

$$\bar{y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{k,i} \tag{3.2}$$

- n is the total number of forecasts,
- d is the number of discrete probability bins,
- n_k is the number of forecasts in the k-th bin,
- $\bar{y}_k$ is the observed relative frequency for the k-th probability bin,
- $\bar{y}$ is the overall observed relative frequency.

[6]https://library.wmo.int/records/item/56227-guidance-on-verification-of-operational-seasonal-climate-forecasts

The term $(\bar{y}_k - \bar{y})^2$ captures the variance between individual forecast categories and the overall event frequency. Higher resolution indicates that forecasts better differentiate between events and non-events.
so the resolution tells us how the model change with different situations.
the scores used to evaluate resolution are Brier Score and Reliability.

3.3.2 Discrimination

Discrimination measures how well the forecast separates cases where the event occurs from cases where it does not. In other words, it examines whether the forecast probabilities differ for events that happen versus those that don't. High discrimination: A forecast has high discrimination if, for example, when rain occurs, the forecast consistently predicts a high probability of rain, and when rain doesn't occur, it predicts a low probability. It means the forecast is good at distinguishing between rain and no-rain days. Low discrimination: If the forecast provides similar probabilities regardless of whether it rains or not (e.g., predicting a 60% chance of rain every day), it has poor discrimination because it doesn't effectively differentiate between days with and without rain. The score used to evaluate descrimination is ROC[7].

3.3.3 Reliability

A forecast is reliable if the predicted probabilities match the actual frequencies. For instance: If you forecast a 40% probability for below-normal rainfall, below-normal rainfall should occur in 40% of the cases where you make that prediction. Similarly, if you forecast a 25% chance of above-normal rainfall, above-normal rainfall should happen 25% of the time when you give that probability. If this relationship holds consistently over many forecasts, the forecasts are well-calibrated (or reliable). A Reliable but Uninformative Forecast A forecast that always gives the climatological probability (e.g., always predicting a 33% chance for each category: below, normal, above normal) would be reliable because the climatological average matches the observed frequencies. However, this forecast wouldn't provide any information about changing conditions from case to case—it doesn't adapt to the current situation, making it uninformative.

$$\textbf{Reliability} = \frac{1}{n} \sum_{k=1}^{d} n_k \left(\bar{p}_k - \bar{y}_k\right)^2 \tag{3.3}$$

- n is the total number of forecasts,
- d is the number of discrete probability bins,
- n_k is the number of forecasts in the k-th bin,
- $\bar{y}_k$ is the observed relative frequency for the k-th probability bin,
- $\bar{p}_k$ is relative frequency for the k-th probability.

3.3.4 Sharpness

Sharp forecasts provide a strong signal about the expected outcome. For example, a sharp forecast might assign a 70% chance to a certain outcome, like above-normal rainfall. This high probability

[7]Relative operating characteristics

communicates more confidence in that specific outcome. On the other hand, when the forecast probabilities are close to the climatological values (like assigning a 40% chance to above-normal, 35% to normal, and 25% to below-normal), the forecast is not very sharp, meaning the forecaster isn't very confident in predicting any one outcome. The climatological probabilities are reliable, but aren't sharp.

3.3.5 The Brier Score (BS)

The Brier Score (BS)[8] is the mean squared differences between pairs of forecast probabilities p and the binary observations y. N is the total forecast number. It measures the total probability error, considering that the observation is 1 if the event occurs, and 0 if the event does not occur (dichotomous events).

$$BS_j = \frac{1}{N}\sum_{i}^{N}(y_{j,i} - p_{j,i})^2$$

where:

- n is the number of forecasts
- $y_{j,i}$ is 1 if the i^th observation was in category j, and is 0 otherwise.
- $p_{j,i}$ is the i^th forecast probability for category j.

The BS takes values in the range of 0 to 1. ***Perfect forecasts receive 0*** and less accurate forecasts receive higher scores. Under the condition that x is 0.5 when the observation data is uncertain, the mean squared differences between the forecast probabilities and observation at 0.5 is calculated.

3.3.6 The ranked probability score (RPS)

The Ranked Probability Score (RPS) is a performance metric used in probabilistic forecasting to assess how well the predicted probability distribution matches the observed outcome distribution. It is particularly useful when there are multiple categories (e.g., terciles such as lower, middle, and upper) and is commonly applied in fields such as meteorology, climatology, and economics.

$$RPS = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{k=1}^{m-1}\left(\sum_{j=1}^{k}(y_{j,i} - p_{j,i})\right)^2$$

where :

- n is the number of forecasts.
- m is the number of categories.
- $y_{j,i}$ is 1 if the i^th observation was in category j, and is 0 otherwise.
- $p_{j,i}$ is the i^th forecast probability for category j

[8] wmo guidance verification

The score is the average squared "error" in the cumulative probabilistic forecasts, and it ranges between 0% for perfect forecasts (a probability of 100% was assigned to the observed category on each forecast) to a maximum of 100% that can only be achieved if all the observations are in the outermost categories, and if the forecasts are perfectly bad (a probability of 100% was assigned to the opposite outermost category to that observed).

3.3.7 Relative operating characteristics

The ROC[9] can be used in forecast verification to measure ***the ability of the forecasts to distinguish an event from a non-event***. For seasonal forecasts with three or more categories, the first problem is to define the "event". One of the categories must be selected as the current category of interest, and an occurrence of this category is known as an event. An observation in any of the other categories is defined as a non-event and no distinction is made as to which of these two categories does occur. So, for example, if below normal is selected as the event, normal and above normal are treated equally as non-events.

the score indicates the probability of successfully discriminating below-normal observations from normal and above-normal observations. It indicates how often the forecast probability for below normal is higher when below normal actually does occur compared to when either normal or above normal occurs.

3.3.8 Relative operating characteristics Skill Score

The Relative Operating Characteristic Skill Score (ROCSS) is a measure used in forecast verification to assess the ability of probabilistic forecasts to discriminate between events and non-events. It builds on the Relative Operating Characteristic (ROC) curve, which plots the hit rate (true positive rate) against the false alarm rate (false positive rate) at various forecast probability thresholds.

- The ROC curve evaluates the discrimination capability of a forecast, i.e., how well the forecast can separate occurrences of an event (e.g., below-normal temperature) from non-events (e.g., normal or above-normal temperature).
- The ROC Skill Score quantifies the area under the ROC curve (AUC) and compares it to a no-skill forecast.

$$ROCSS = \frac{AUC - AUC_{no-skill}}{1 - AUC_{no-skill}}$$

where:

- AUC : Area Under the ROC Curve for the forecast being evaluated.
- $AUC_{no-skill}$: Area Under the Curve for a no-skill forecast 0.5 for our case.

Interpretation of ROCSS:

- 1: Perfect discrimination ability.
- 0: No skill (forecast performs no better than random guessing).
- Negative values: Forecast performs worse than random guessing.

3.3.9 summary

[9] wmo guidance verification

Metric	Focus	What it Measures	Dependent on Observed Outcomes?	Visualization/Tools
Reliability	Probabilities match observed frequencies	Calibration of probabilities	Yes	Reliability diagram
Discrimination	Differentiating between outcomes	Ability to distinguish events from non-events	Yes	ROC curve, AUC
Sharpness	Boldness of probabilities (away from average)	Confidence of the forecast	No	Histogram of forecast probabilities
Resolution	Informativeness and variability of forecast	Ability to provide specific, useful info	Yes	Brier Score decomposition

Table 3.1: Key differences between reliability, discrimination, sharpness, and resolution in seasonal forecasting.

CHAPTER 4

RESULTS

This chapter presents the results of our study, divided into two main sections: temperatures and precipitation. In each section, we provide a comprehensive analysis of the deterministic and probabilistic evaluation of forecast performance across the MENA region. By examining both temperature and precipitation metrics, we aim to highlight the strengths and limitations of the forecasting models, offering valuable insights into their reliability and applicability in this climatically diverse area.

4.1 Temperature

In the temperature session, the use of heatmaps and temperature metrics maps will allow for a visual interpretation of model performance across various metrics. By analyzing these heatmaps, we can identify the most effective models based on metrics such as Spearman rank correlation, RMSE (Root Mean Square Error), and the Coefficient of Determination (R^2). These visualizations will help in understanding the relationships between predicted and observed temperatures, aiding in the selection of models with the highest predictive accuracy. Additionally, the use of probabilistic evaluation metrics such as the Brier Score, Reliability, Ranked Probability Score (RPS), and Relative Operating Characteristics (ROC) will provide insights into the model's performance in terms of forecast quality, focusing on calibration, discrimination, and sharpness. These methods will be used to refine and improve predictive models for temperature forecasting in the MENA region.

4.1.1 Deterministic evaluation results

Correlation

The provided images used in this section displays the correlation between observed and modeled surface temperatures across the MENA region for four different seasons: June, July, and August (JJA); December, January, February (DJF); March, April, May (MAM); and September, October, November (SON). Each map visualizes the predictive accuracy of various climate models, with the color intensity representing the strength and direction of the correlation between the models' predictions and the observed data. Hatched areas on the maps indicate regions where the correlation is statistically significant ($p \leq 0.05$). These maps are essential for evaluating the performance of climate models across different seasons and understanding which models provide the best temperature forecasts for the MENA region.

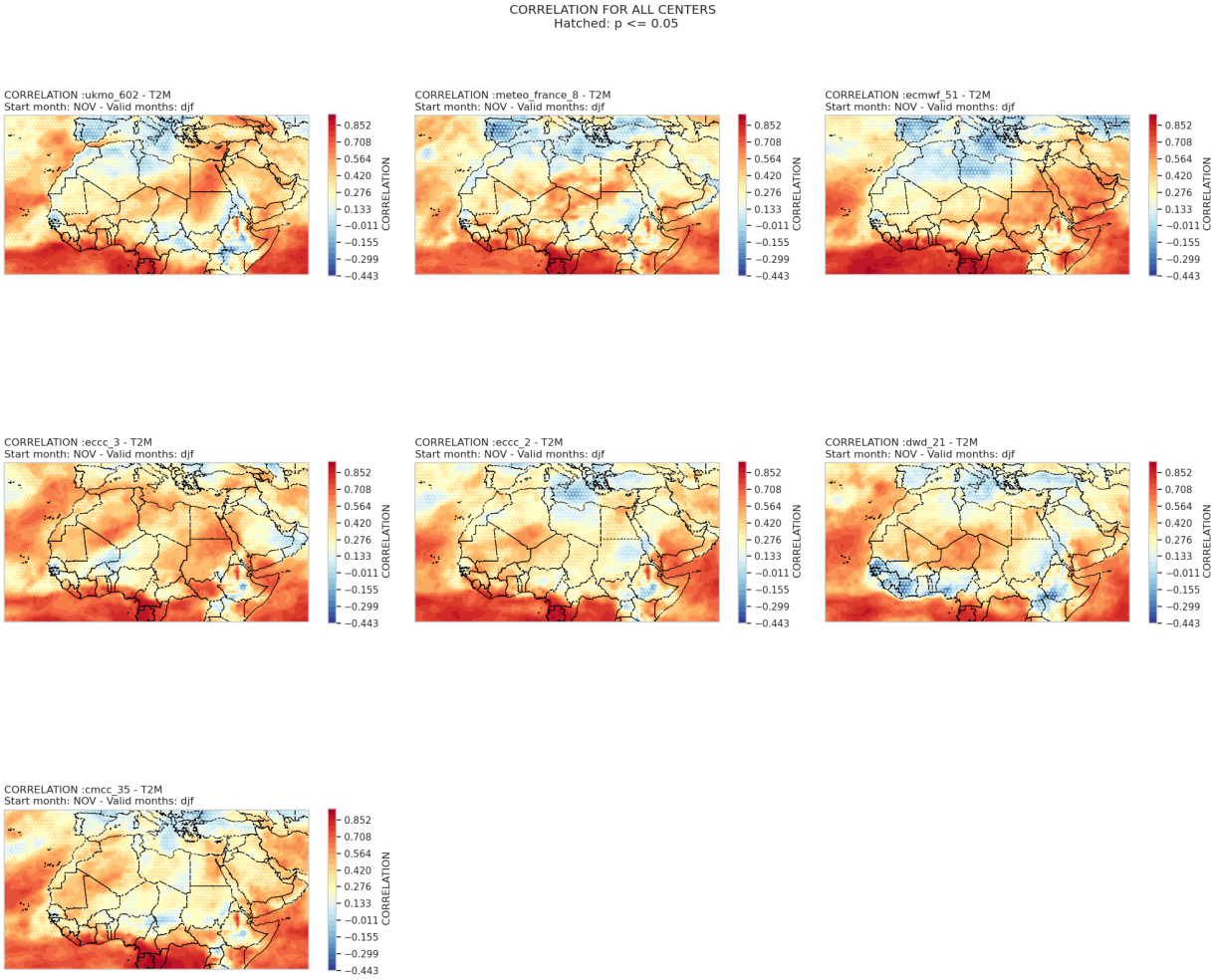

Figure 4.1: Temperature correlation maps for DJF (December-January-February)

The analysis of the **Spearman rank correlation for djf** from the figures above shows the following trends:

- For the **UKMO model**, the correlation between observed and modeled temperatures is high (around 0.7) in southeastern Libya, Egypt, and northwestern Sudan. However, it is low and negative in Morocco, Algeria, and parts of Libya.
- The **Meteo-France model** exhibits a similar pattern, with high correlations in southeastern Libya and Egypt but low and negative values in Morocco and Algeria.
- In contrast, the **ECMWF model** shows higher correlations in Saudi Arabia and Yemen.
- For the **ECCC System 3 model**, the correlations are notably higher over a large part of the region, indicating better agreement between observed and modeled temperatures.

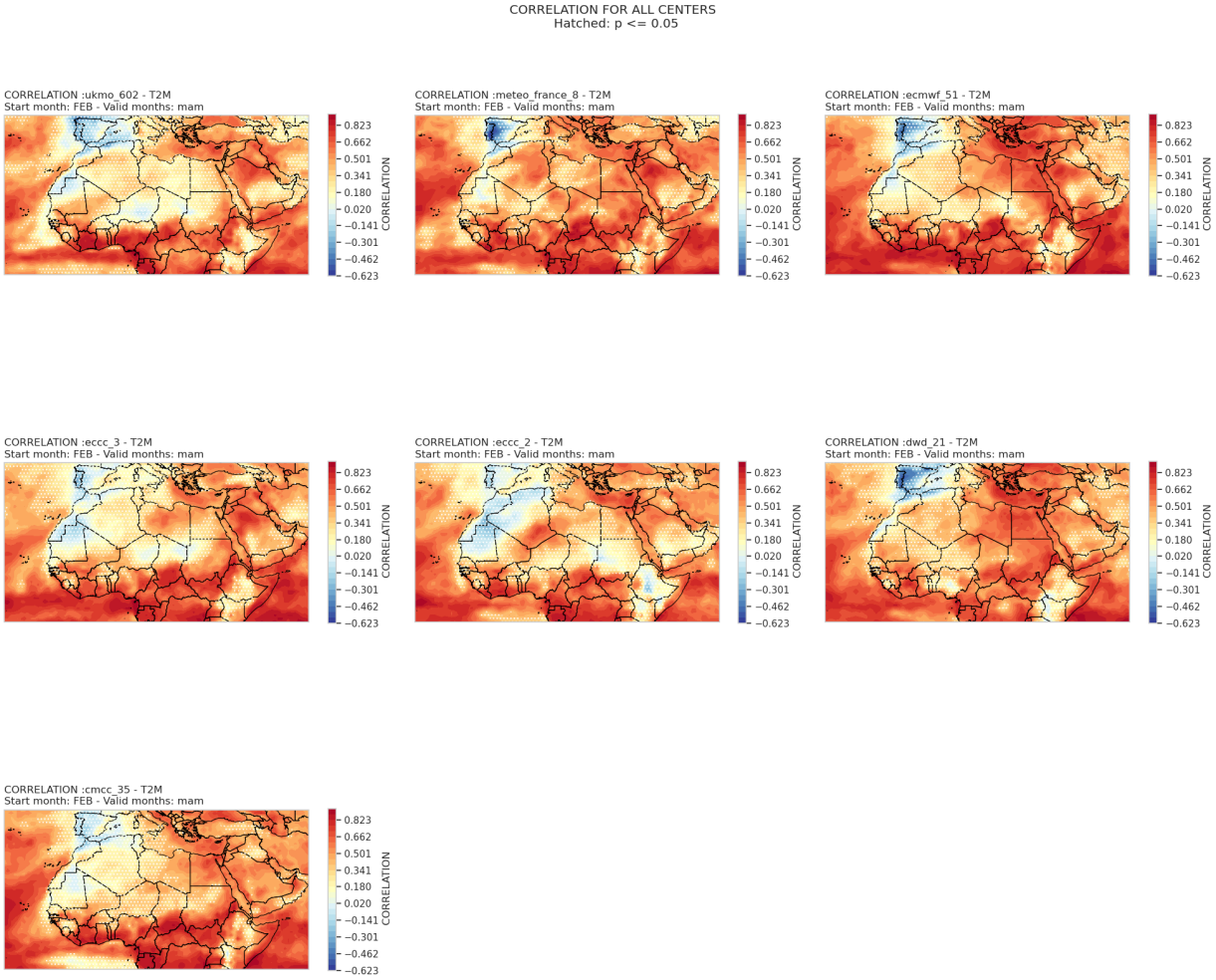

Figure 4.2: temperature correlation maps for mam

For MMA, the Met Office, ECMWF, and DWD demonstrate better performance across this region, except in northern Morocco, where correlations range between approximately -0.12 and 0.02.

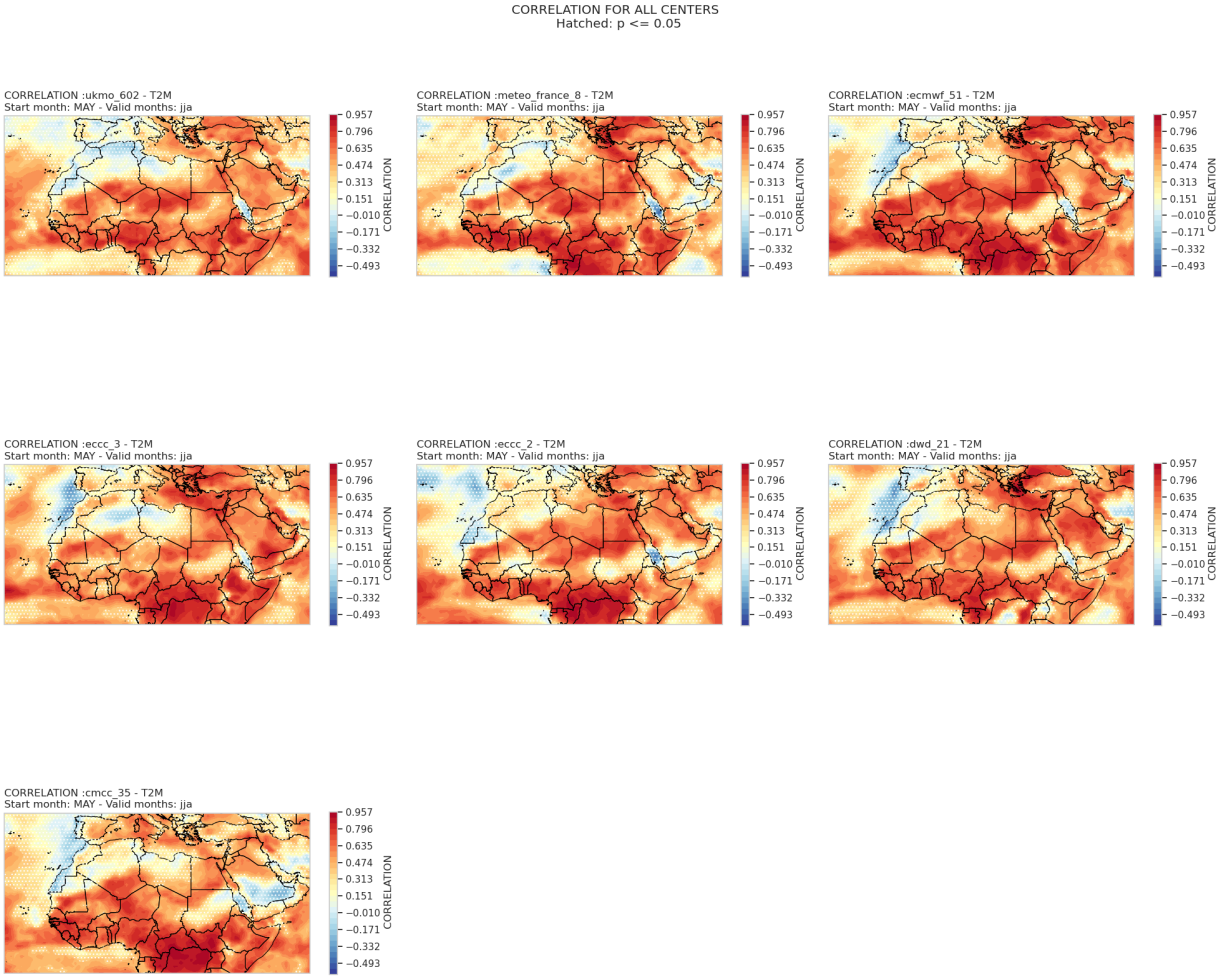

Figure 4.3: temperature correlation maps for jja

For JJA, the ECMWF and ECCC-2 models exhibit the best correlations overall, except in southern Morocco and Tunisia, where their performance is notably weaker.

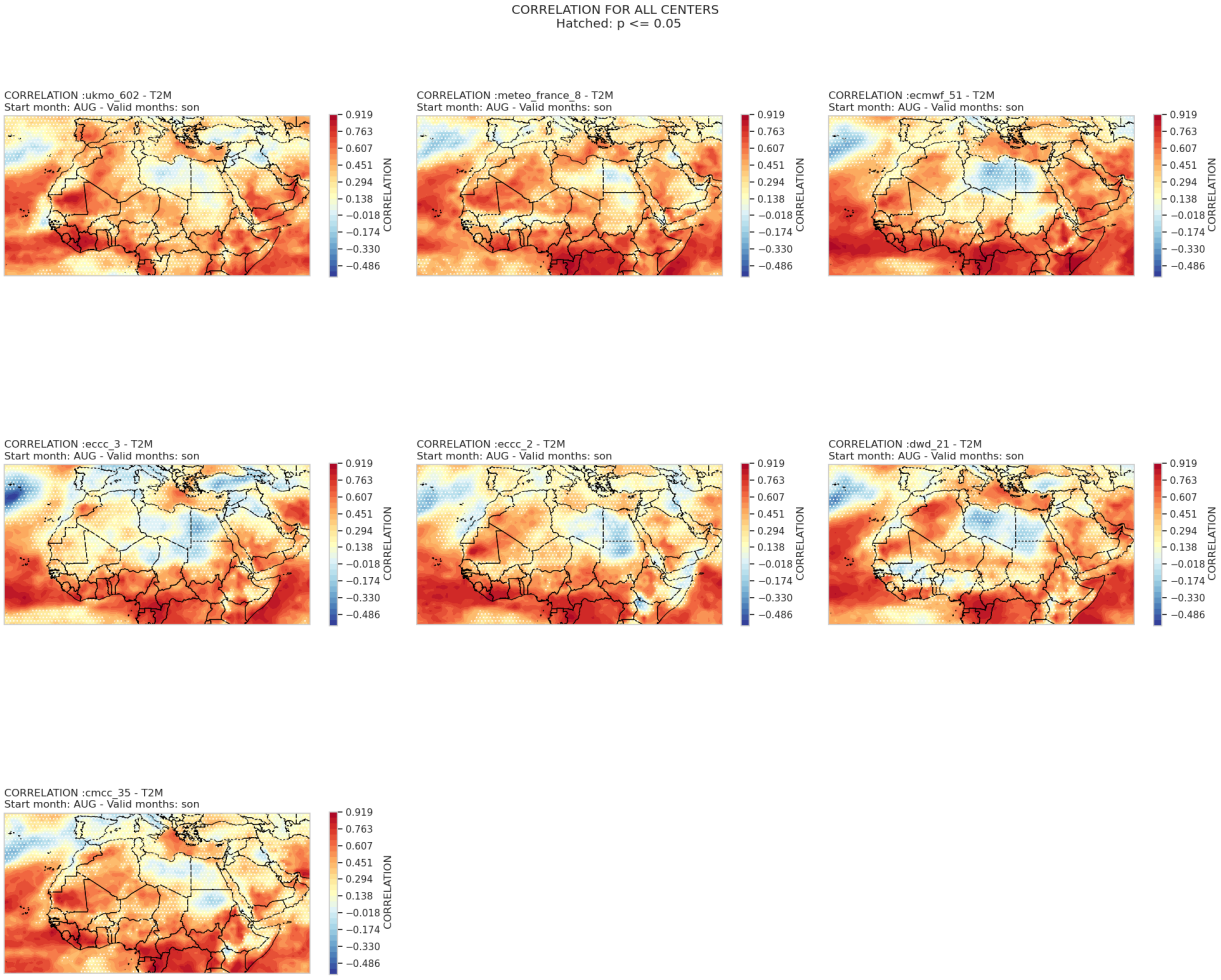

Figure 4.4: temperature correlation maps for son

For SON , the Meteo-France model performs well, except in Egypt, showing better correlations in Morocco compared to other models and seasons. Similarly, the ECMWF model also performs well but shows weaker correlations in Libya and Egypt.

Figure 4.5: Temperature correlation heatmap for mena region

The heatmaps provide an average overview of the models' performance across the MENA region for a given season and a forecast lead time ranging from 1 to 3 months. In general, correlations tend to decrease as the lead time increases from 1 to 3 months. For instance, the ECMWF model shows higher performance during JJA, whereas the ECCC-2 model performs better during MAM. As previously mentioned, these results represent averages over the entire region. For a more detailed analysis, a mask was applied to focus specifically on North African countries within the MENA region. The results of this targeted analysis are shown in the heatmap below.

Figure 4.6: Temperature correlation heatmap for north africa

Root Mean Square Error

The maps in this section show the RMSE (Root Mean Square Error) between observed and modeled surface temperatures across the MENA region for the four seasons: JJA, DJF, MAM, and SON. The RMSE, expressed in the same units as temperature, evaluates the accuracy of climate models, with lower values indicating better performance. These maps help identify the strengths and limitations of the models across seasons, contributing to the improvement of climate forecasts for the region.

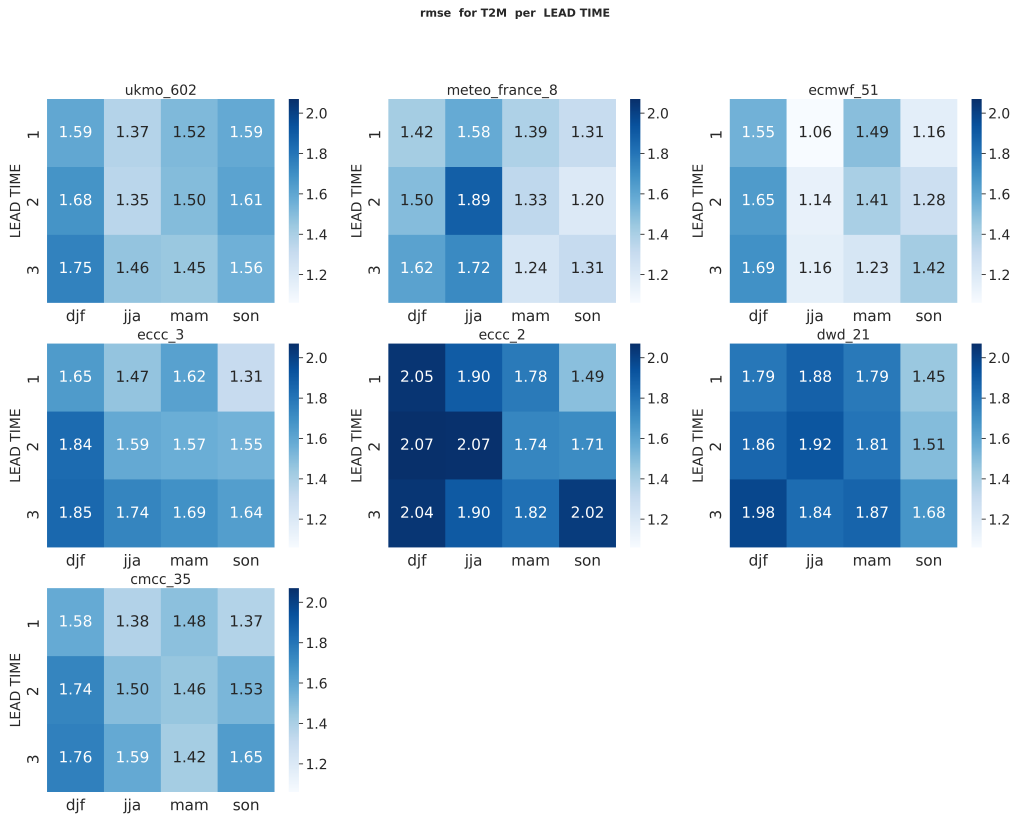

Figure 4.7: Temperature rmse heatmaps for all the sasons

The heatmap for the seven models highlights distinct variations in their seasonal performance. The UKMO model shows moderate to good performance, particularly in JJA and MAM, as reflected by relatively low RMSE values ranging between 1.35°C and 1.51°C across the three lead times. This indicates that the UKMO model is reasonably effective in capturing surface temperature variations during these seasons, likely benefiting from its ability to simulate key atmospheric processes during these periods.

In contrast, Météo-France exhibits weaker performance in JJA, with higher RMSE values suggesting less accurate predictions of surface temperatures during this season. This could be attributed to the model's limitations in capturing summer-specific temperature drivers in the MENA region, such as heatwaves, desert-air interactions, or seasonal atmospheric circulation patterns.

The ECMWF model emerges as the best-performing model based on the heatmap, particularly during JJA, where it demonstrates high predictive accuracy and consistency. This is supported by the RMSE map, which shows significantly lower RMSE values across most of the MENA region. These low RMSE values confirm the ECMWF model's ability to capture regional temperature dynamics effectively, with reduced errors across diverse climatic zones.

Notably, the spatial distribution of RMSE on the map reinforces the ECMWF's strong performance, as it maintains relatively low error values in critical parts of the MENA region. This suggests that the ECMWF model is better equipped to account for the complex climatic interactions in the region, such as the influence of desert regions, coastal temperature gradients, and

seasonal weather patterns.

Overall, these findings emphasize the importance of selecting climate models based on their seasonal and spatial performance, as they play a critical role in improving the accuracy of temperature forecasts for the MENA region.

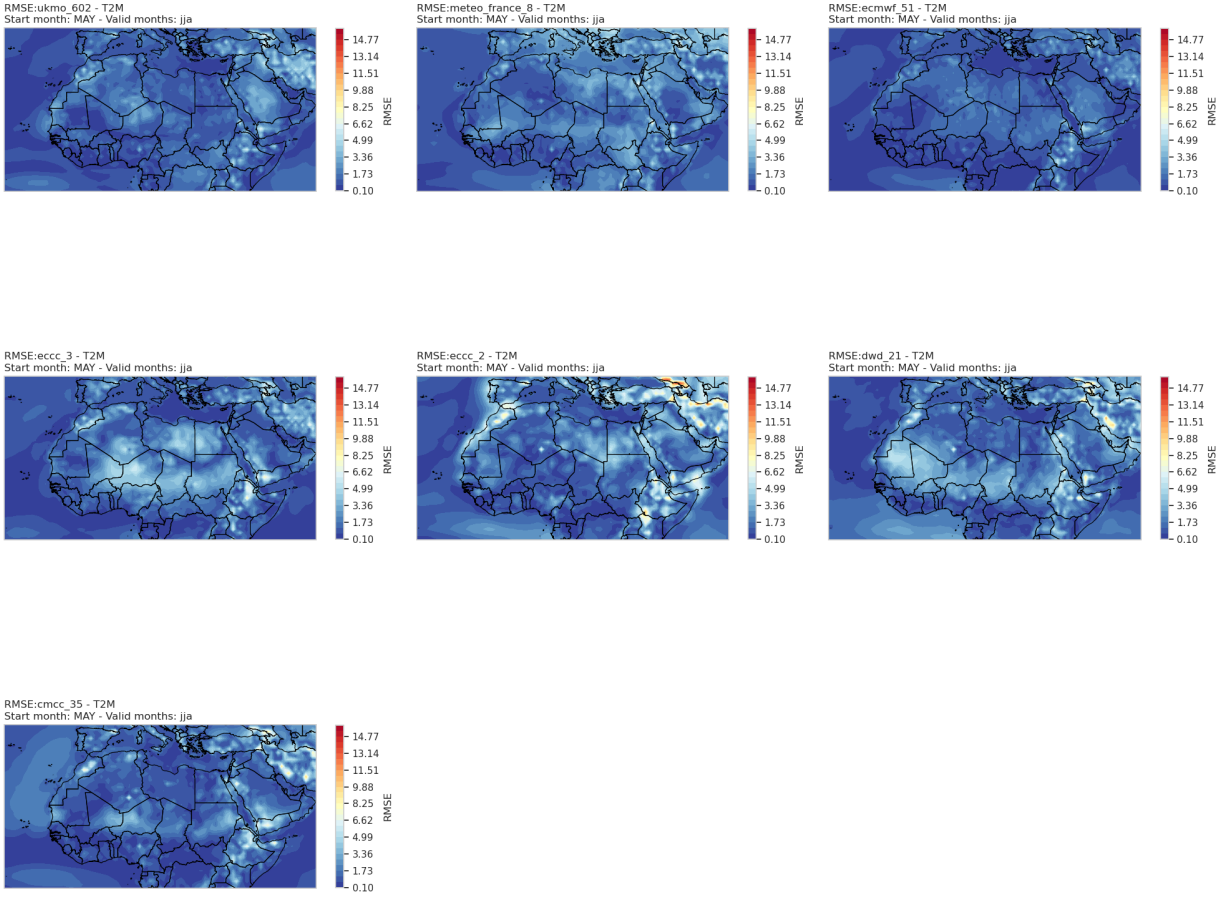

Figure 4.8: temperature rmse maps for jja

To better understand the performance of climate models in sub-regions of MENA, a focus on North Africa is essential due to its distinct climatic features, such as vast desert landscapes, high seasonal temperature variability, and localized weather phenomena. The heatmap provides a comparative analysis of RMSE values for the seven models across this region, highlighting their seasonal accuracy in predicting surface temperatures.

The North African climate, heavily influenced by the Sahara Desert and Mediterranean coast, presents unique challenges for climate modeling, such as accurately simulating extreme temperatures and their spatial distribution. By examining the heatmap, we can identify which models

perform well in capturing these specific climatic patterns and assess their suitability for seasonal forecasting in this critical sub-region of MENA.

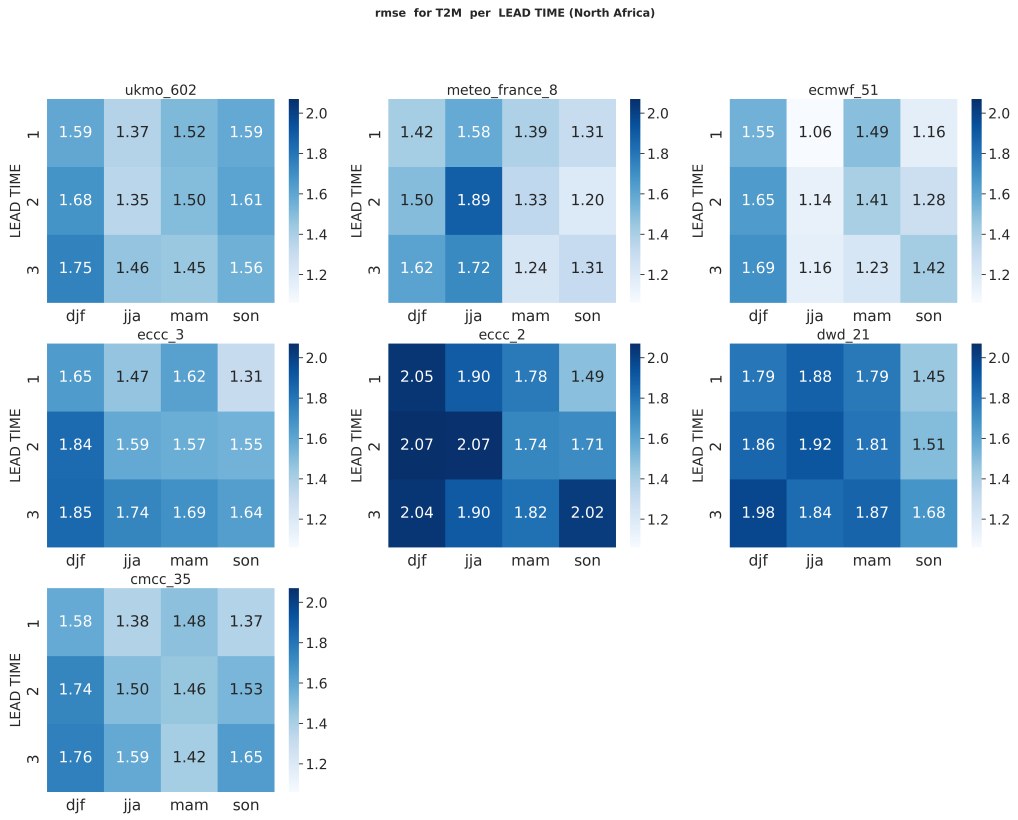

Figure 4.9: Temperature rmse heatmap for north africa

This specific analysis revealed that, although the ECMWF model performs well across the MENA region, its performance in North Africa is comparatively lower, except during JJA. As highlighted earlier, JJA remains the most accurate season for ECMWF in this sub-region, with RMSE values ranging between 1.34°C and 1.58°C. This indicates that the model effectively captures temperature dynamics in the summer, likely due to its ability to simulate key seasonal factors such as strong solar heating and atmospheric circulation patterns specific to North Africa during this period.

In contrast, Météo-France demonstrates consistent performance across the three lead times and, on average, performs well over the four seasons in North Africa, particularly during SON. This could be attributed to the model's ability to account for transitional weather conditions in autumn, such as changes in atmospheric pressure systems and coastal influences that significantly impact temperatures in this region.

These observations underline the importance of analyzing sub-regional performance to identify models that are better suited for specific areas and seasons. While ECMWF excels in JJA, Météo-France shows a more balanced performance throughout the year, making it potentially more reliable for multi-seasonal forecasting in North Africa.

Coefficient of Determination (R^2)

The maps in this section show the R-SQUARED between observed and modeled surface temperatures across the MENA region for the four seasons: JJA, DJF, MAM, and SON. R-SQUARED is a statistical measure that indicates how well the model explains the variability in observed data, with values closer to 1 signifying better performance. These maps provide valuable insights into the predictive skill of the climate models, highlighting their ability to capture seasonal temperature patterns.

Figure 4.10: Temperature rsquared heatmaps for all the seasons

Based on this deterministic metric (R-SQUARED), the ECMWF model demonstrates superior performance for lead time 1 across all four seasons, particularly during MAM. In general, the portion of variance explained by the model decreases as the lead time increases. This indicates that while the model is highly effective at capturing seasonal variability of surface temperatures in the short term, its predictive skill diminishes over longer time horizons.

The strong performance during MAM highlights the ECMWF ability to capture the complexities of spring, a season marked by transitional weather patterns in the MENA region. The high R-SQUARED values during this period suggest that the model accurately reflects observed temperature variability by effectively simulating key drivers such as the gradual warming trend, atmospheric circulation changes, and the interaction between desert and coastal dynamics.

Such precision underscores the ECMWF model's reliability for short-term seasonal forecasting, particularly during periods of heightened climatic variability like MAM. However, the decreasing performance with increasing lead times suggests the need for careful interpretation of forecasts beyond lead time 1, as uncertainty increases with longer projections.

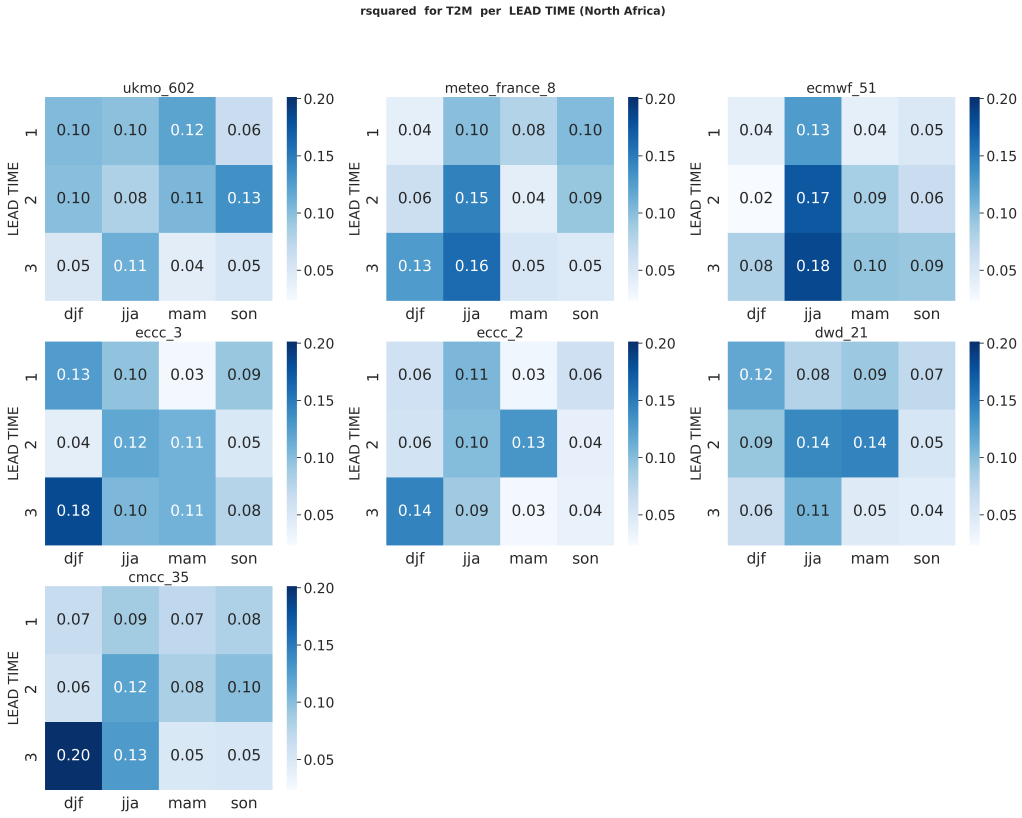

Figure 4.11: Temperature rsquared heatmaps for all the seasons North Africa

A closer look at North Africa reveals that the ECMWF model performs best during JJA, with R-SQUARED values increasing as the lead time increases-contrary to the trend observed for the MENA region as a whole, where performance typically decreases with longer lead times. This suggests that the ECMWF model is particularly adept at capturing the persistent summer temperature patterns in North Africa, which may benefit from stronger model predictability over time due to the relatively stable atmospheric and climatic conditions during JJA.

Similarly, Météo-France also shows good performance in North Africa, maintaining consistent R-SQUARED values across different lead times and seasons. This consistency highlights the model's ability to handle the diverse climatic features of the region, such as the extreme temperatures influenced by the Sahara Desert and the moderating effects of the Mediterranean coastline.

These findings emphasize the importance of regional analysis, as the performance of climate models can vary significantly within sub-regions of MENA. While ECMWF excels in predicting North Africa's summer temperatures, the observed increase in R-SQUARED with lead time underscores the need to investigate the underlying factors driving this unusual trend, which contrasts

with the broader MENA region's dynamics.

4.1.2 Probabilistic evaluation results

To complement the deterministic evaluation of model performance, probabilistic evaluation metrics are employed to assess the reliability and skill of climate models in predicting the likelihood of specific outcomes. Unlike deterministic metrics, which focus on the accuracy of single-point predictions, probabilistic metrics evaluate the quality of the models' forecast distributions, accounting for uncertainty and variability in predictions. These metrics are essential for understanding how well models represent the range of possible outcomes, particularly in regions like MENA, where climatic variability and extremes are prominent. By incorporating probabilistic metrics, this analysis provides a more comprehensive evaluation of the models' predictive capabilities and their usefulness in decision-making under uncertainty. The figures illustrate two main approaches to probabilistic assessment metrics, including the Brier Score (BS) and others. The first approach averages across lead times and grid points, while preserving categories, where the final figure contains the value of the metric for each season across all four seasons, for each category (mean, lower, upper) defined by the 1/3 quartiles. This method provides insight into the predictive ability of models under different seasonal conditions and forecast probability categories, particularly how well models capture temperature variations in the middle, lower, and upper quartiles of the predicted probability distribution. The diversity of metric values across these categories helps highlight the sensitivity of models to different levels of forecast confidence. It indicates their ability to differentiate between forecast uncertainty and actual observed outcomes, providing a nuanced understanding of how accurately models predict different temperature ranges.

The second approach averages all grid points in the MENA region while retaining all lead times and seasons. This aggregated view provides an overall assessment of the measure for each season, considering all lead times and forecast categories. This approach focuses on how models perform across different forecast scenarios and how well they produce accurate and reliable temperature forecasts, regardless of forecast probability. By retaining lead times and seasons in the analysis, this method provides a comprehensive picture of model performance over time and under different climate conditions. It reveals how well models generalize across different forecast scenarios, helping to identify which models are most effective at producing consistent and reliable forecasts.

Brier score

The figure at the bottom illustrates that most models demonstrate relatively high performance, as indicated by a small Brier Score (BS), meaning that the predicted probabilities are close to the observed ones. This reflects accurate forecast probabilities forT2M The UKMO model, however, shows moderate performance with a larger BS, suggesting that its predicted probabilities are less closely aligned with observed outcomes compared to other models. Notably, the middle category presents lower performance relative to the other two categories (lower and upper). This indicates that while some models capture temperature variability well in extreme conditions (upper category), their skill diminishes when forecasting moderate changes (middle category). This discrepancy highlights the challenges models face in translating predicted probabilities into reliable forecasts, particularly for temperature variations that are neither extreme nor outlier events.

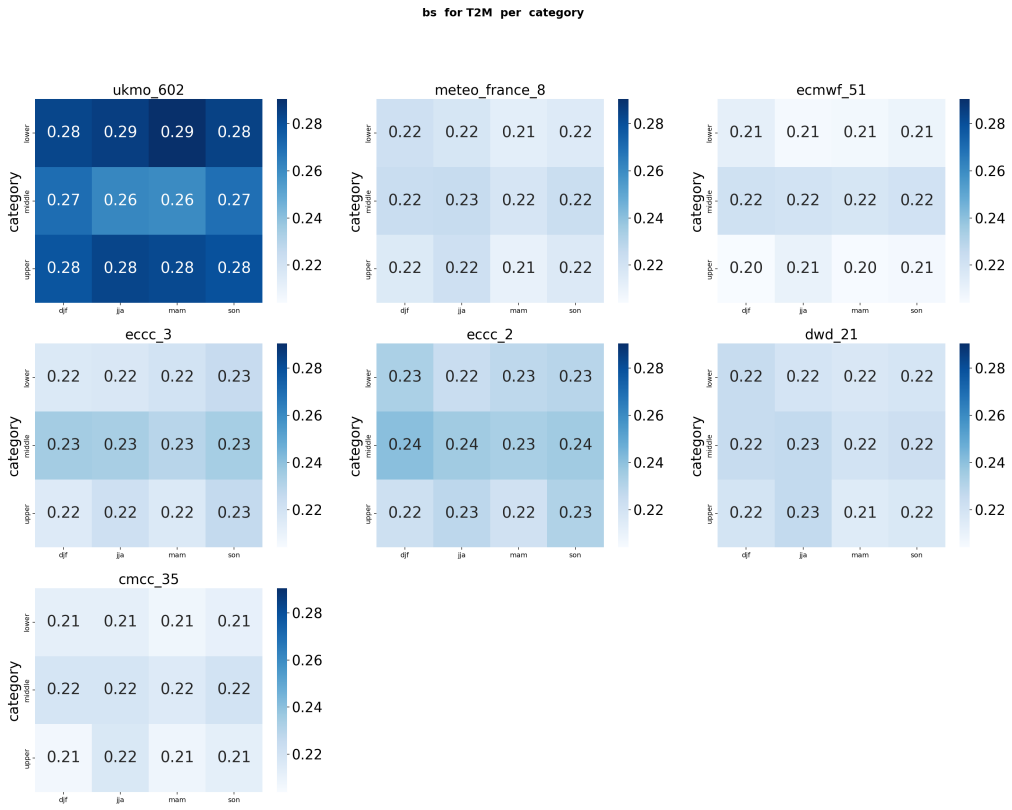

Figure 4.12: Temperature Brier score heatmaps for all the seasons per categories

An analysis by lead time revealed that the models Meteo France, ECMWF, and CMCC exhibit superior performance, as indicated by lower Brier Scores (BS). This suggests that these models provide more accurate probabilistic forecasts for T2M compared to others. Moreover, the differences in BS values between successive lead times are minimal, indicating that the predictive skill of these models remains relatively consistent as the forecast horizon increases.

This stability in performance across lead times is particularly noteworthy, as it reflects the robustness of these models in maintaining their ability to produce reliable forecasts over time. The lower BS values also suggest that these models effectively capture the relationship between predicted probabilities and observed outcomes, ensuring high confidence in their probabilistic predictions. Such consistent performance across lead times is crucial for operational forecasting, as it highlights these models' reliability for both short- and medium-term forecasts in the MENA region.

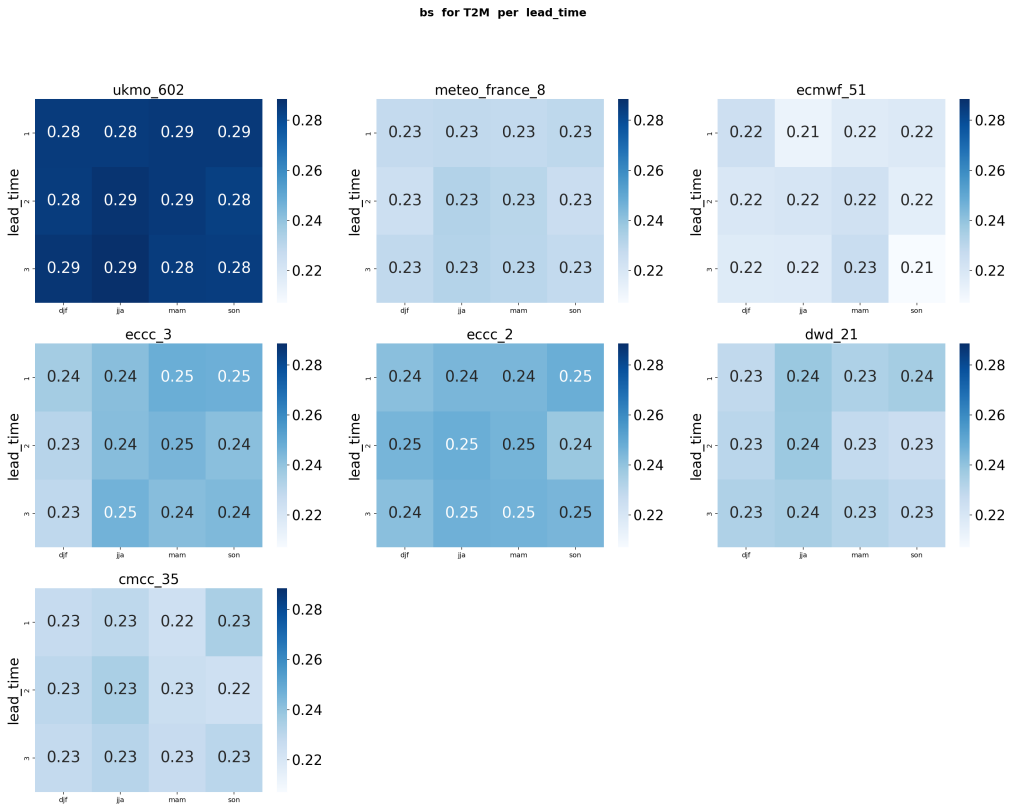

Figure 4.13: Temperature brier score heatmaps for all the seasons per lead time

o further evaluate the performance of these models and assess whether the same interpretations apply, the analysis will now focus specifically on the North African region. By examining the Brier Score metrics in this sub-region, we aim to determine whether the observed trends, such as the consistency of model performance across lead times and the relative strengths of ECMWF, CMCC, and Meteo France, remain valid or if different patterns emerge due to the unique climatic conditions and variability in North Africa.

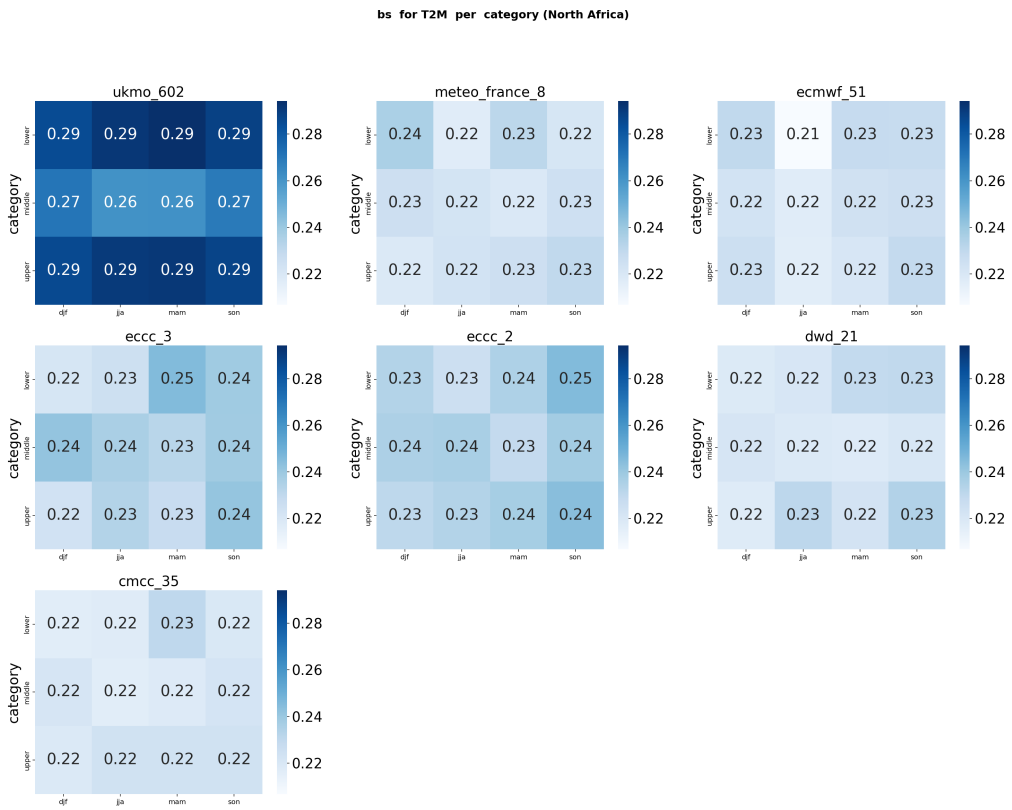

Figure 4.14: Temperature Brier score heatmaps for all the seasons per categories for north africa

A closer analysis focused on the North African region revealed that the performance of the top-performing models—ECMWF, CMCC, and Meteo France—remains consistent. The zoomed evaluation shows that these models continue to exhibit low Brier Scores, indicating reliable probabilistic predictions across lead times. This finding suggests that the unique climatic conditions of North Africa do not significantly alter the predictive skill of these models, reaffirming their robustness and adaptability to sub-regional variations within the broader MENA context.

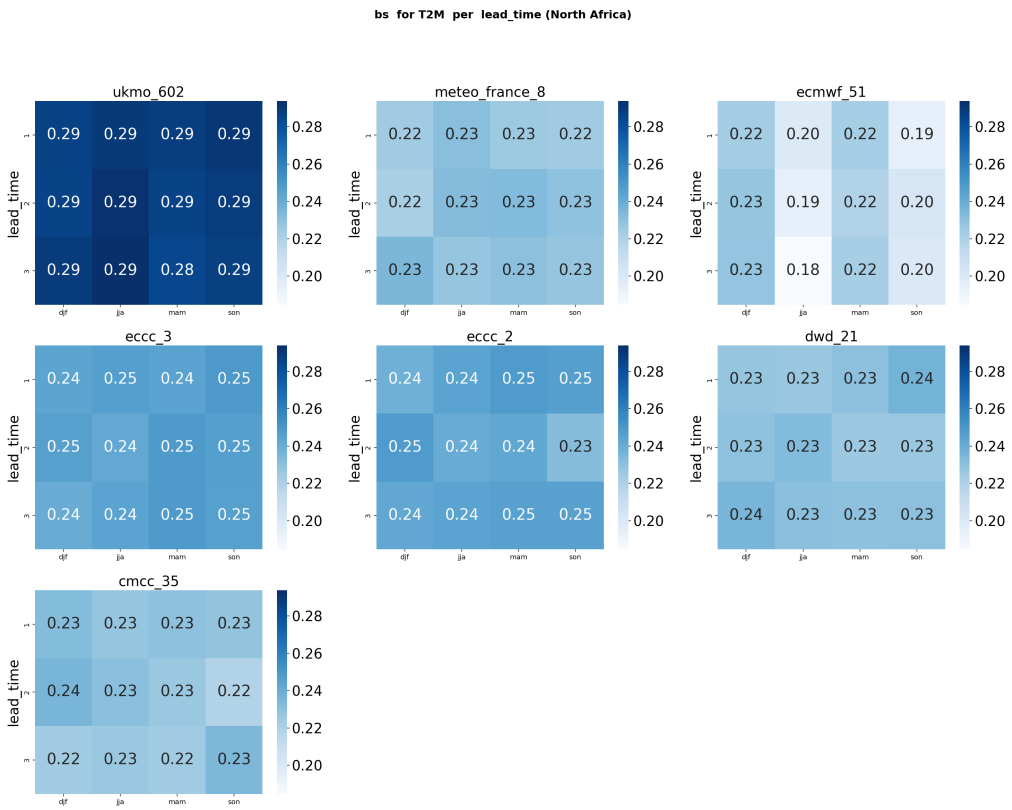

Figure 4.15: Temperature brier score heatmaps for all the seasons per lead time North Africa

The results by lead time reveal similar conclusions, showing a consistent stability in Brier Scores across different lead times. This indicates that the predictive skill of the top-performing models, such as ECMWF, CMCC, and Meteo France, remains robust even as the forecast horizon increases. The minimal variation in scores across lead times underscores the reliability of these models for producing accurate probabilistic forecasts, regardless of the temporal range considered.

Reliability

The reliability diagrams presented for the season assess the probabilistic performance of the seven climate models across three forecast categories: lower tercile , middle tercile , and upper tercile . These diagrams compare the predicted forecast probabilities to the observed frequencies, providing a clear indication of the models' ability to produce reliable temperature forecasts. A perfectly reliable model would align closely with the diagonal, where predicted probabilities match observed outcomes. By analyzing these diagrams, we can identify which models perform best for specific categories and uncover any systematic biases, such as over- or under-forecasting certain probabilities. This evaluation is crucial for understanding the strengths and weaknesses of the models, particularly in their ability to capture temperature extremes and moderate conditions .

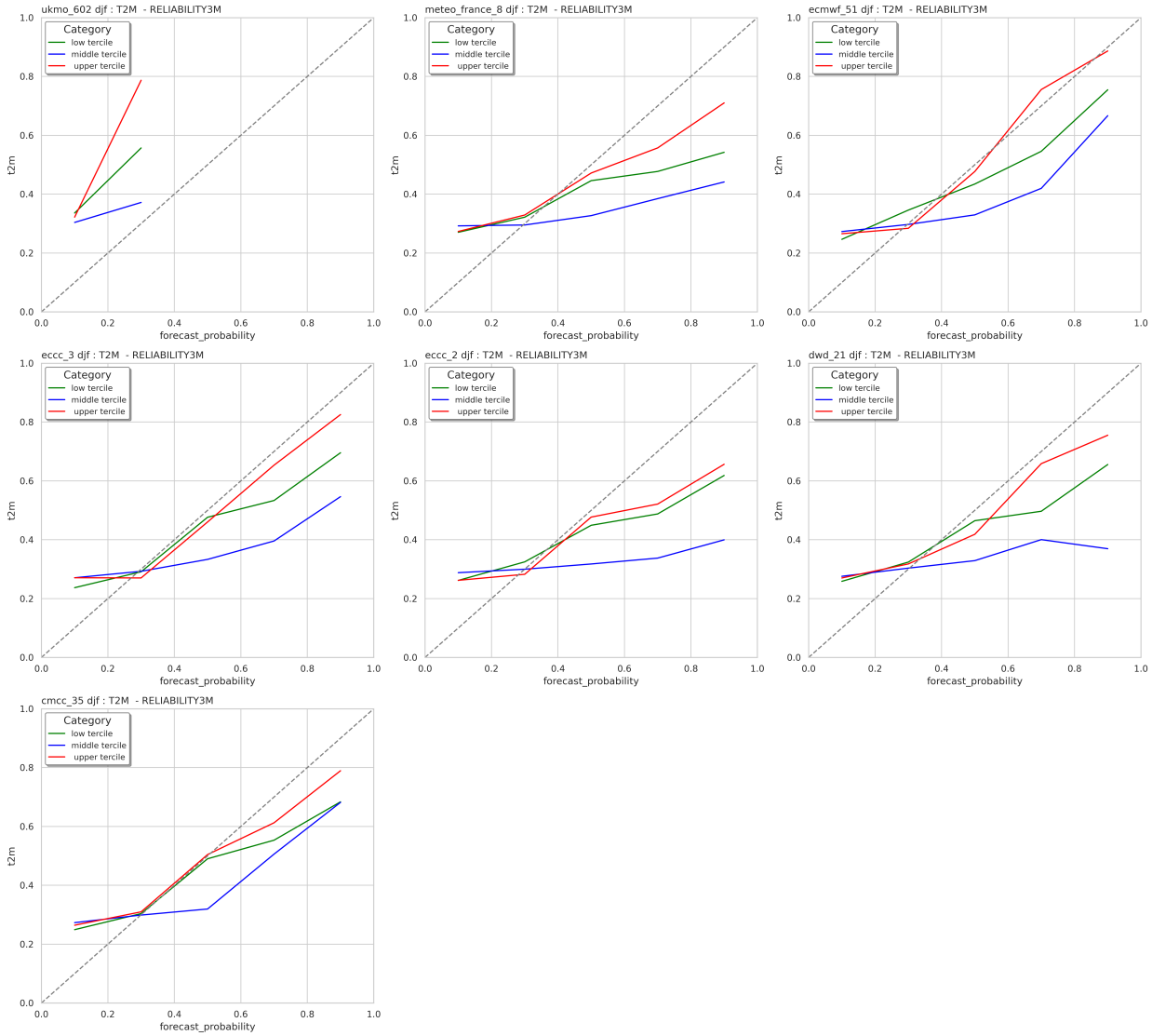

Figure 4.16: temperature reliabilty maps for djf

For the DJF season, the reliability diagrams show that most models exhibit similar performance, with the exception of UKMO, which significantly overestimates probabilities across all categories. For ECMWF, there is a noticeable overestimation in the upper tercile (warm extremes) starting at a forecast probability of approximately 0.6, indicating that the model tends to assign higher probabilities to warm events compared to their observed frequencies. Conversely, ECMWF shows a systematic underestimation for both the lower and middle terciles across most probabilities. For the remaining models, except UKMO, there is a general trend of underestimation across all three categories, beginning at forecast probabilities around 0.5. This pattern suggests that while the models capture some aspects of the observed temperature distribution, they struggle to provide accurate probabilities for more extreme or moderate conditions, especially at higher forecast probabilities. These results highlight the challenges faced by the models in reliably predicting temperature outcomes during the winter season (DJF), particularly for extreme categories. For the other three seasons (JJA, MAM, and SON), the reliability diagrams reveal a general tendency for most mod-

els to underestimate probabilities across all tercile categories, starting at approximately 0.4. This underestimation indicates that the models tend to be over-cautious, predicting lower probabilities than what is observed in reality. However, UKMO still stands out as an exception, showing a consistent overestimation across the forecast range, particularly for higher probabilities. This suggests that UKMO tends to assign excessively high probabilities to events that occur less frequently than predicted, which reduces its reliability. Overall, this recurring underestimation in the majority of models highlights their difficulty in accurately capturing the likelihood of temperature outcomes, especially beyond moderate probabilities.

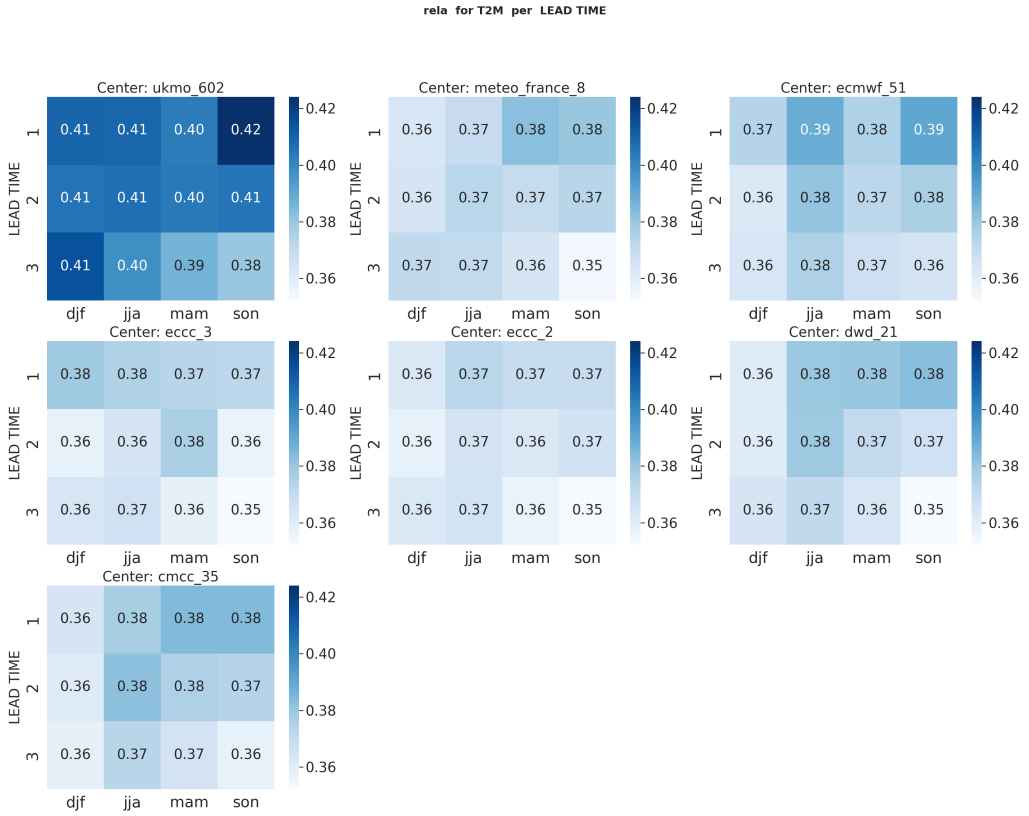

Figure 4.17: temperature reliabilty heatmap

The heatmap further confirms the analyses derived from the reliability diagrams. It highlights similar trends observed across the models and seasons.Conversely, UKMO continues to stand out with its consistent overestimation of probabilities, aligning with the overconfidence seen in the reliability diagrams. This agreement between the heatmap and the reliability curves strengthens confidence in the identified patterns .

Figure 4.18: temperature reliabilty heatmap for north africa

A more focused analysis on North Africa has not significantly altered the overall conclusions derived from the broader MENA region. The consistent performance patterns observed across the different models and seasons remain largely unchanged when examining the North African context.

The ranked probability score

The Ranked Probability Score (RPS) provides a valuable measure of forecast performance by evaluating the accuracy of probabilistic predictions across different categories. It combines both the skill in predicting the occurrence of events and the sharpness of the forecast distribution. By comparing the forecasted probability distribution against the observed outcomes, the RPS quantifies the deviation between the predicted and actual probabilities. A lower RPS value indicates better forecast accuracy, reflecting both how well the forecast aligns with observed frequencies and how well it discriminates between different probability categories. This metric helps to identify which models offer the most reliable probabilistic predictions, particularly in terms of capturing the likelihood of various temperature outcomes within a given forecast period.

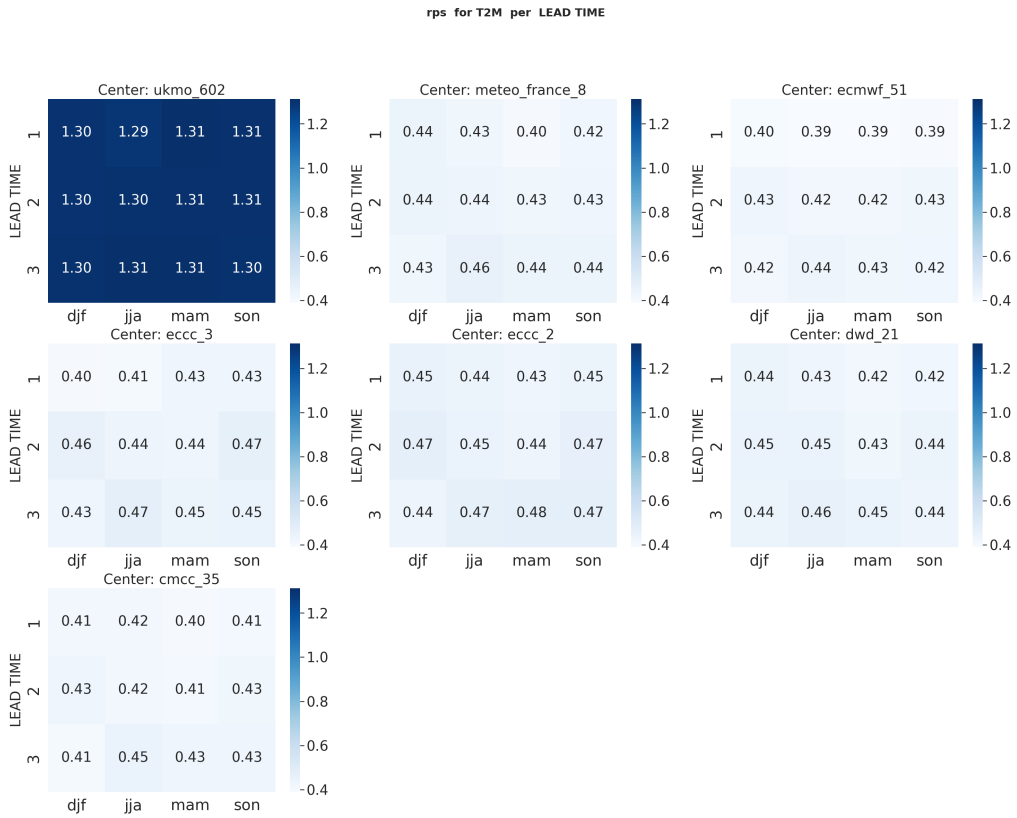

Figure 4.19: Temperature RPS heatmaps for all the seasons per categories

The figure displaying the Ranked Probability Score (RPS) for different climate models and seasonal periods provides a detailed view of model performance across various start months (DJF, JJA, MAM, SON). Each cell in the matrix represents the RPS value for a specific model and season combination, with the color intensity indicating how well the forecast probabilities match the observed data.

From this figure, it is evident that ECMWF consistently shows lower RPS values, indicating better predictive accuracy across different seasons. This suggests that ECMWF's forecasts are more closely aligned with observed temperature variability. The relatively higher RPS values for UKMO model underscore their challenges in accurately capturing temperature variations. For the North African region, the results mirror those observed in the broader MENA region.

This suggests that despite the localized focus on North Africa, the model performance differences remain significant, particularly for UKMO.

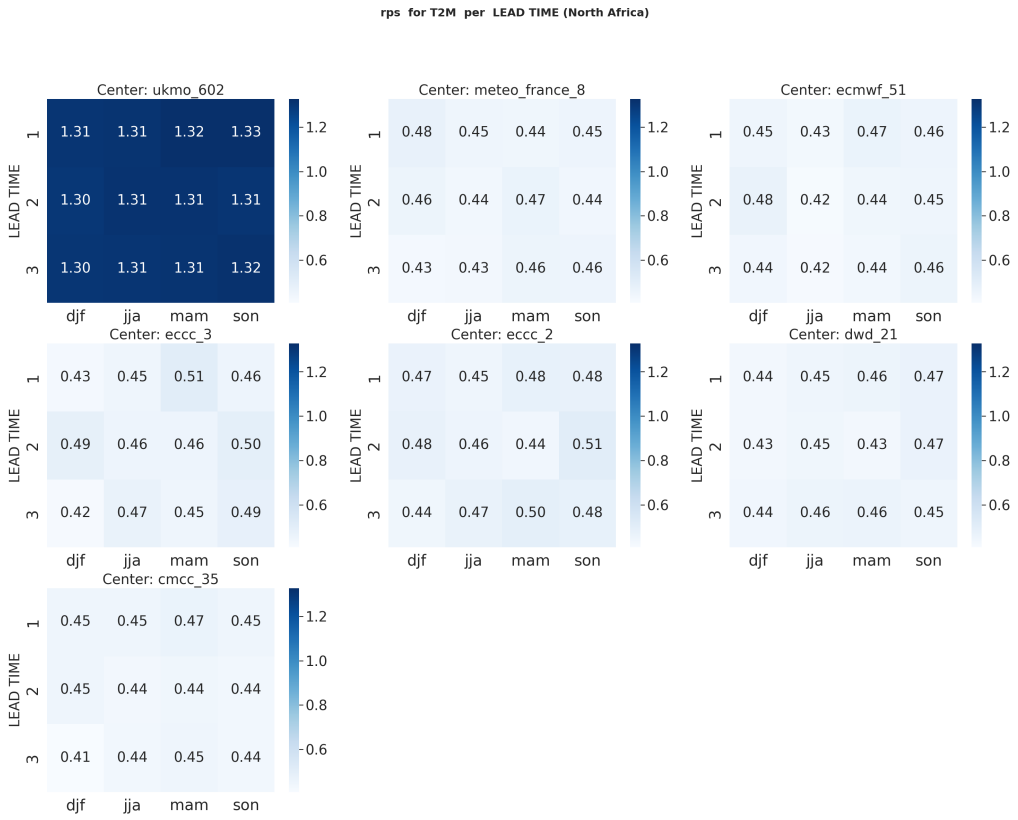

Figure 4.20: Temperature RPS heatmaps for north africa

Receiver Operating Characteristic

The ROC (Receiver Operating Characteristic) curve is an important tool for evaluating the performance of predictive models, particularly in the context of probabilistic forecasts. It provides a graphical representation of the trade-off between the true positive rate and the false positive rate across various threshold levels.

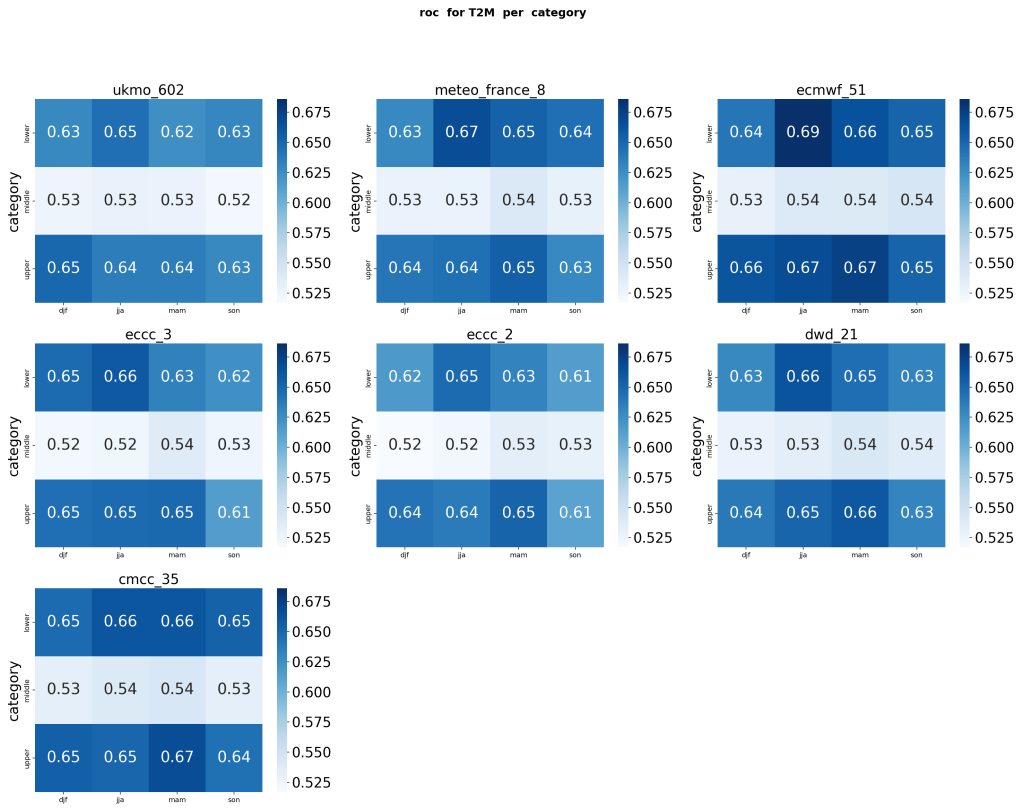

Figure 4.21: Temperature AUC heatmaps

Models generally exhibit similar performance, as indicated by the high Area Under the ROC Curve (AUC) values, which reflect their ability to effectively discriminate between predicted probabilities and observed outcomes. Unlike the reliability metric, where UKMO showed weaker performance, it performs relatively well in terms of the AUC, demonstrating good skill in distinguishing between forecasted events and non-events. Similar to the findings with the Brier Score (BS), the "middle" probability category tends to show weaker performance compared to the "lower" and "upper" categories. This highlights the models' greater sensitivity in accurately predicting events with extreme probabilities (high or low), but reduced skill for moderate probability scenarios. This consistency across metrics underscores the need to address forecast performance specifically in the middle category to further improve model predictions.

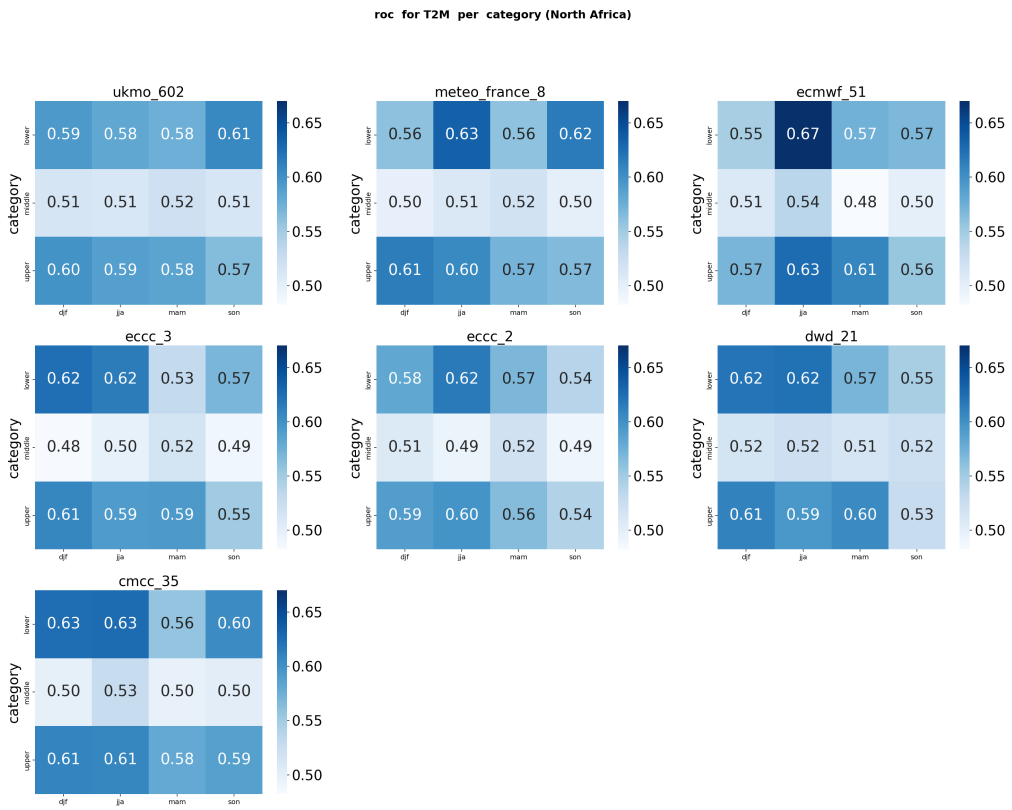

Figure 4.22: Temperature AUC heatmaps for north africa

The figure above confirms the same conclusions for the North Africa region. The models generally maintain similar performance, with high AUC values reflecting strong discrimination skill across all categories. UKMO continues to show robust results in terms of ROC, despite its weaker reliability performance. As observed previously, the "middle" probability category remains the least performant compared to the "lower" and "upper" categories, indicating the models' reduced ability to predict moderate probability events. This consistency in findings suggests that the regional focus on North Africa does not significantly alter the overall assessment of model performance.

Relative operating characteristics Skill Score

ROCSS provides an assessment of a model's ability to discriminate between observed and forecasted events relative to a reference model, often a climatological or random forecast. A higher ROCSS indicates that the model has skill in distinguishing between occurrences and non-occurrences of an event, while a score close to zero suggests no significant improvement over the reference

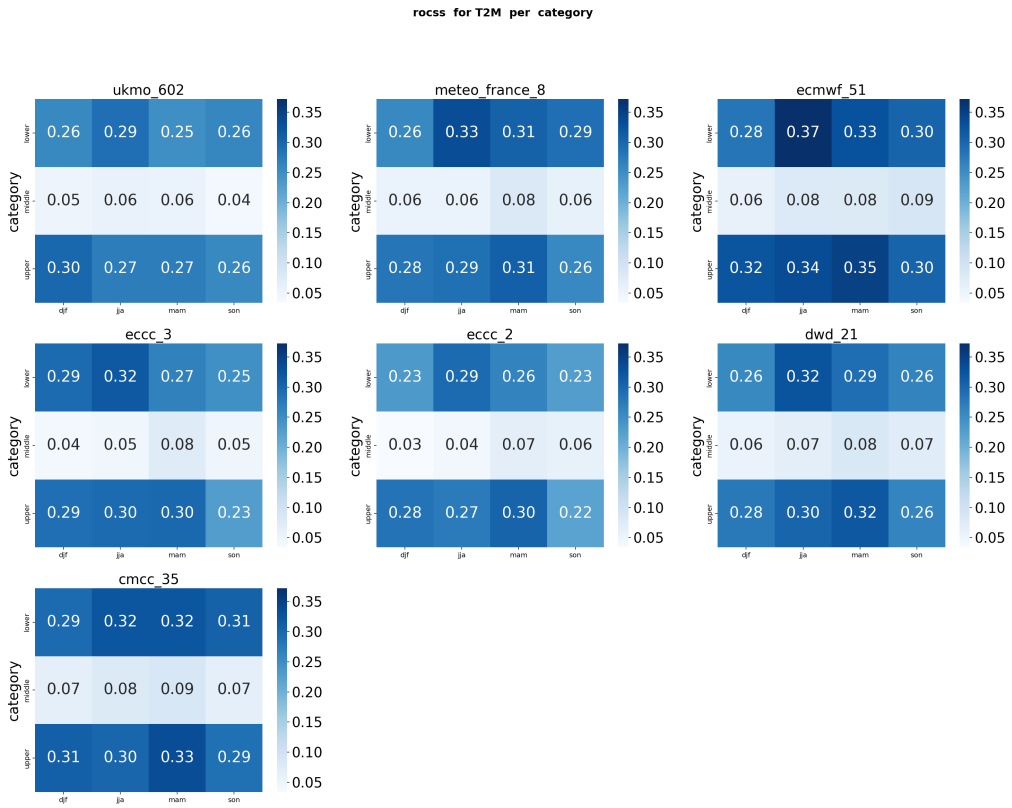

Figure 4.23: Temperature ROCSS heatmaps for MENA region

. The ROC Skill Score (ROCSS) analysis reveals similar conclusions to the ROC results for Mena region and north Africa. The models generally demonstrate consistent and positive skill, highlighting their ability to discriminate between observed and forecasted events. UKMO, which showed weaker performance in reliability metrics, continues to perform well in terms of ROCSS, confirming its relative robustness in event discrimination. Additionally, as observed with the ROC scores, the "middle" category exhibits lower performance compared to the "lower" and "upper" categories. This indicates that while the models excel at predicting extreme events with high or low probabilities, their ability to capture moderate probability events remains limited.

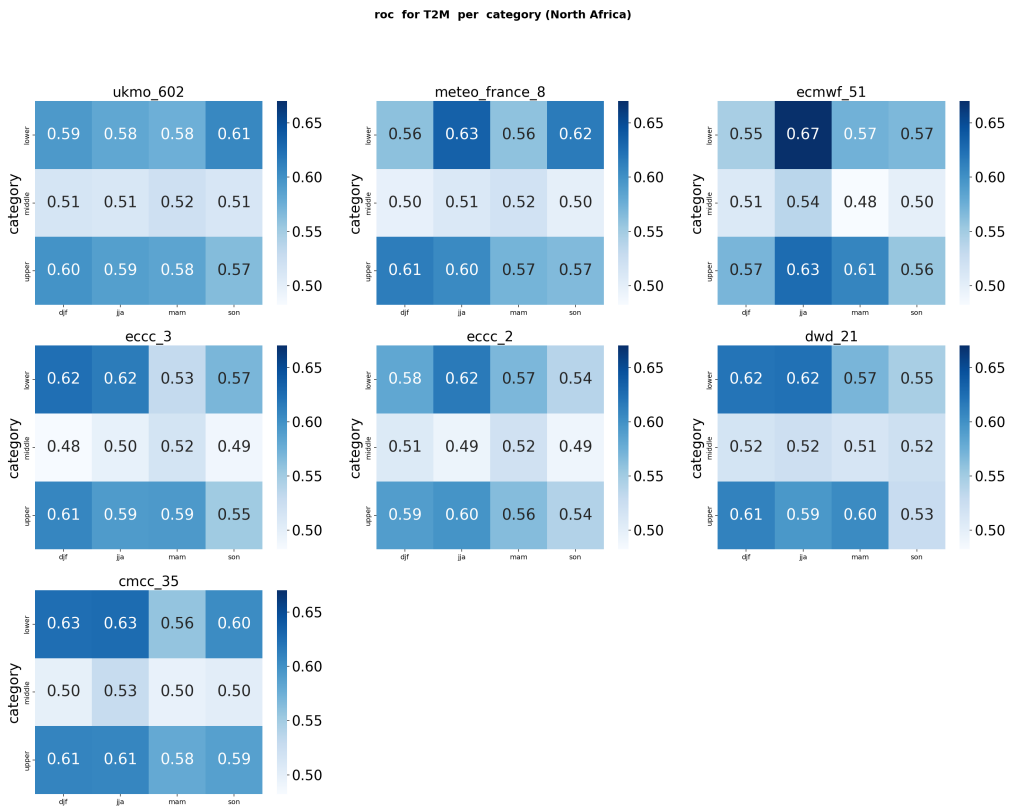

Figure 4.24: Temperature ROCSS heatmaps for north africa

4.2 PRECIPITATIONS

IN general, the forecast of precipitations is more complicated than temperature, thus the scores are a little less good for this part especially the deterministic ones.

4.2.1 Deterministic Evaluation Metrics

Spearman rank correlation

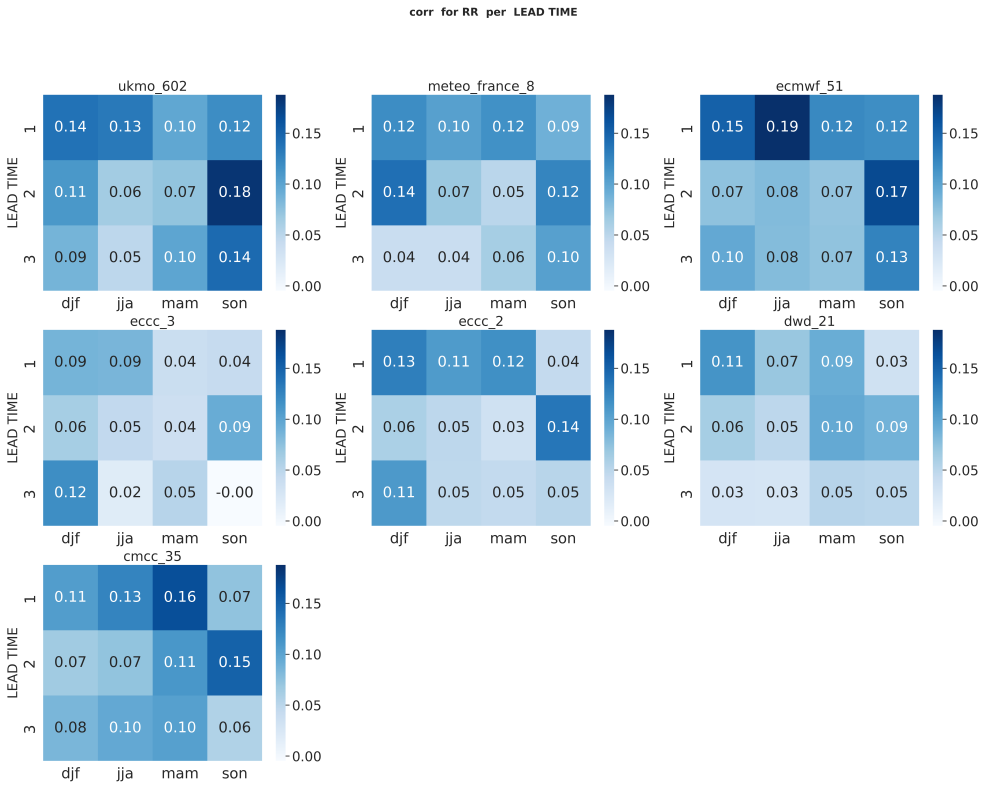

Figure 4.25: The Heatmap of correlation for the mena region for every period ***(1 for perfect Correlation)***

The correlation is weak for all centers; however, the best models are ***ECMWF, UKMO, and CMCC-35***. There is no clear variability in performance along lead-time. For SON, the performance is excellent at lead-time 2 for all centers. As for the other seasons, the performance is generally strong at the 1st lead-time but decreases with increasing lead-time. Hence, Meteo-France also shows good performance, but it decreases significantly over time.

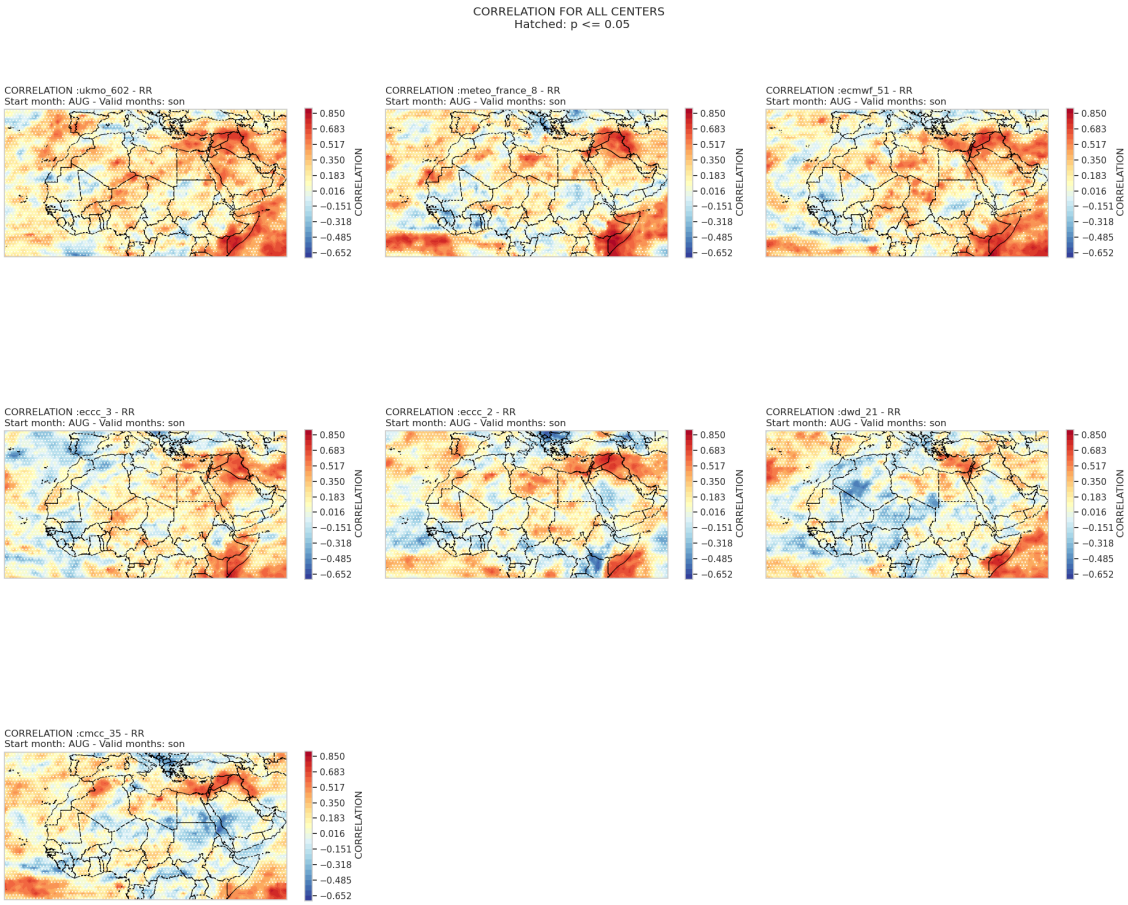

Figure 4.26: 3-months Rolling mean of Spearman Correlation in MENA Region for all centers SON

For temperature, the models demonstrate the best performance in the tropical regions. However, for precipitation, the situation is different. Hence the results show good performance during SON, where the Middle East, East Africa, and North Africa exhibit the highest correlation performance.

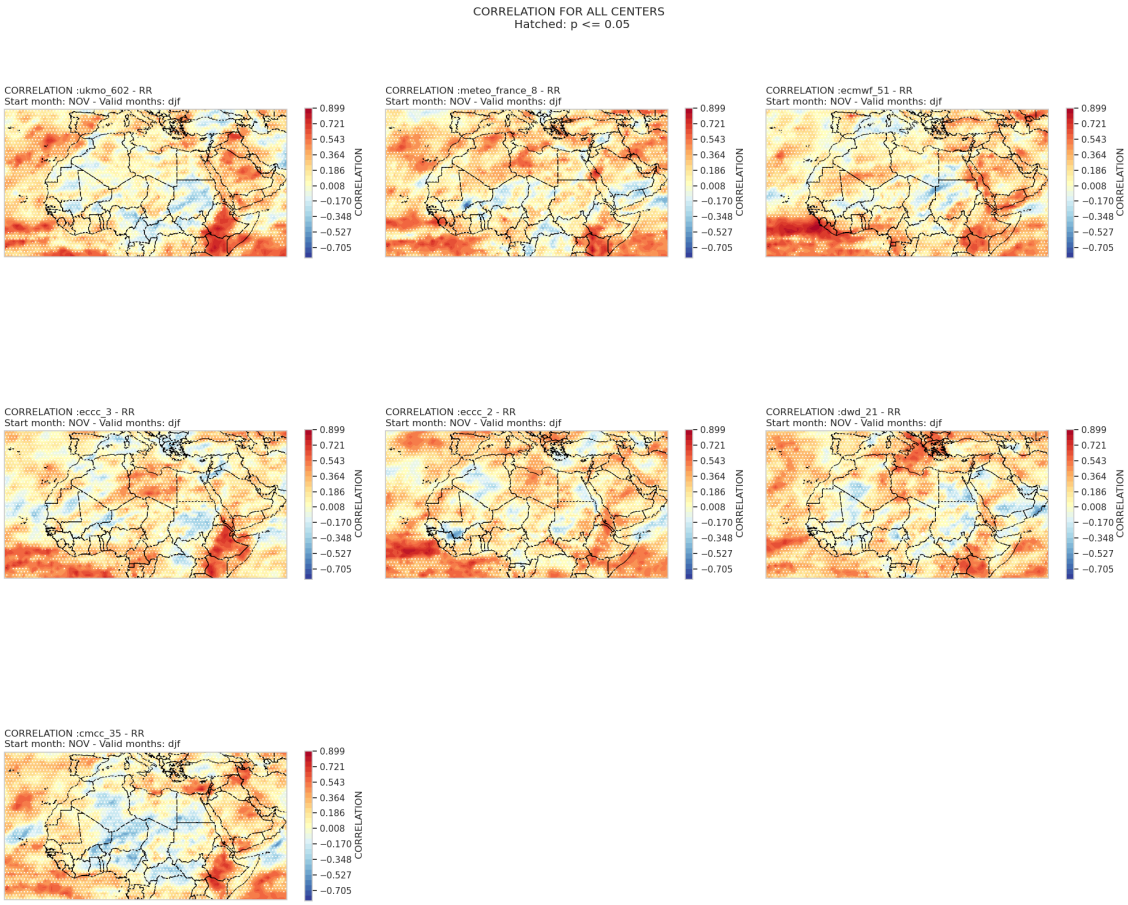

Figure 4.27: 3-months Rolling mean of Spearman Correlation in MENA Region for all centers DJF

The 3-month rolling mean for SON correlation shows that the best models are ***ECMWF, UKMO, and Meteo-France***. The correlation is significant across most of the MENA region, except in the east of Africa, Palestine, Syria, Jordan, and Iraq, where the correlation is maximal, for all centers the Middle East and East of Africa are have the highest score. However, near the equator, the correlation is negative and weak. This results are confirmed in all centers.

For DJF, the situation is generally better than for SON. The best model for North Africa is ***Meteo-France***, as it shows good and significant correlation. In general, ***ECMWF and Meteo-France*** are the best. In general we notice that there is differences between centers, especially in the Middle East and Center Africa.

focus on north africa: according to the heatmap below, the correlation shows no big difference for the first lead-time, but for the second and third lead-times, it became lower. Thus, the ***ecmwf,ukmo and meteo-france*** maintain relatively good correlation.

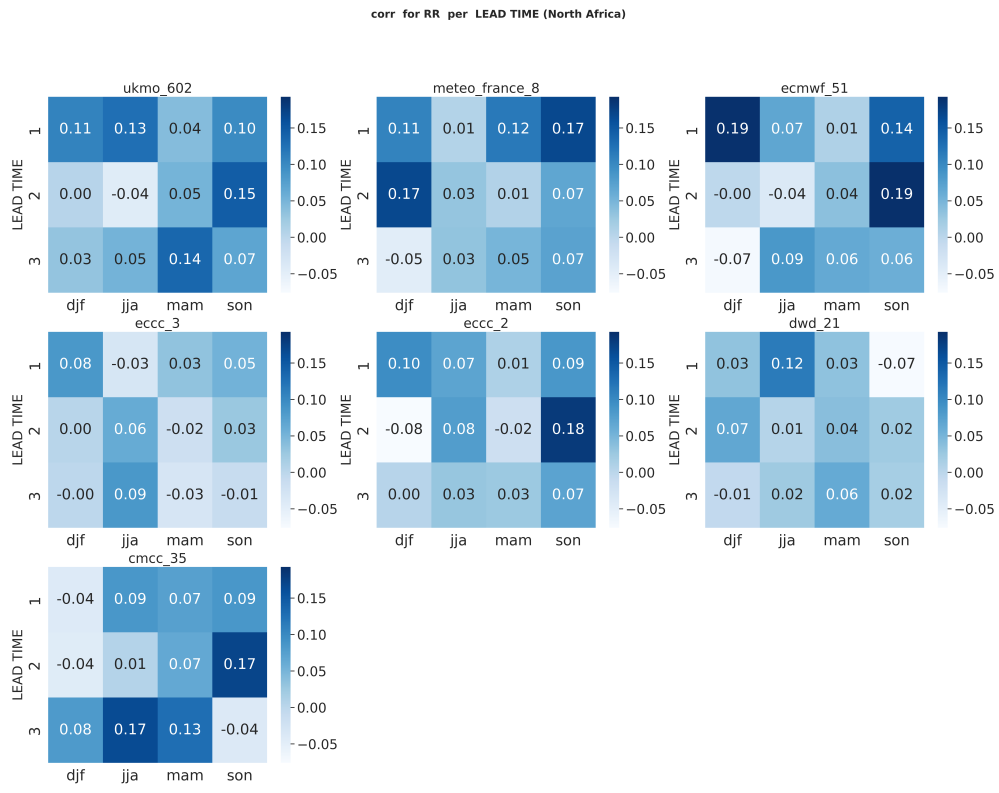

Figure 4.28: The Heatmap of correlation for the mena region for every period ***(1 for perfect Correlation)***

RMSE

for the Root Mean Squared Error, the best models shown in the heatmap below are ***DWD, ECMWF and UKMO***. The RMSE score demonstrate a moderate performance for all models especially ***DWD, ECMWF and UKMO***. The performance is stable over lead-times and it is much better for djf in all centers.

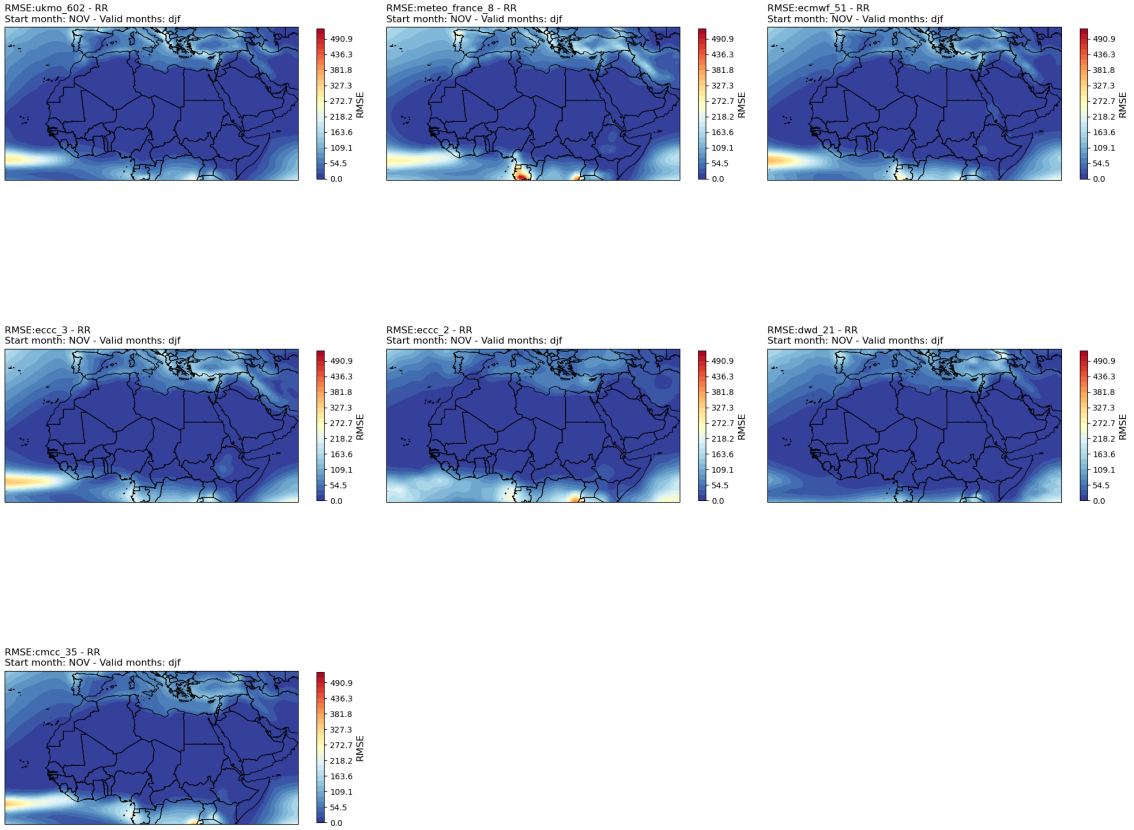

Figure 4.29: 3-months Rolling mean of RMSE in MENA Region for all centers DJF in mm

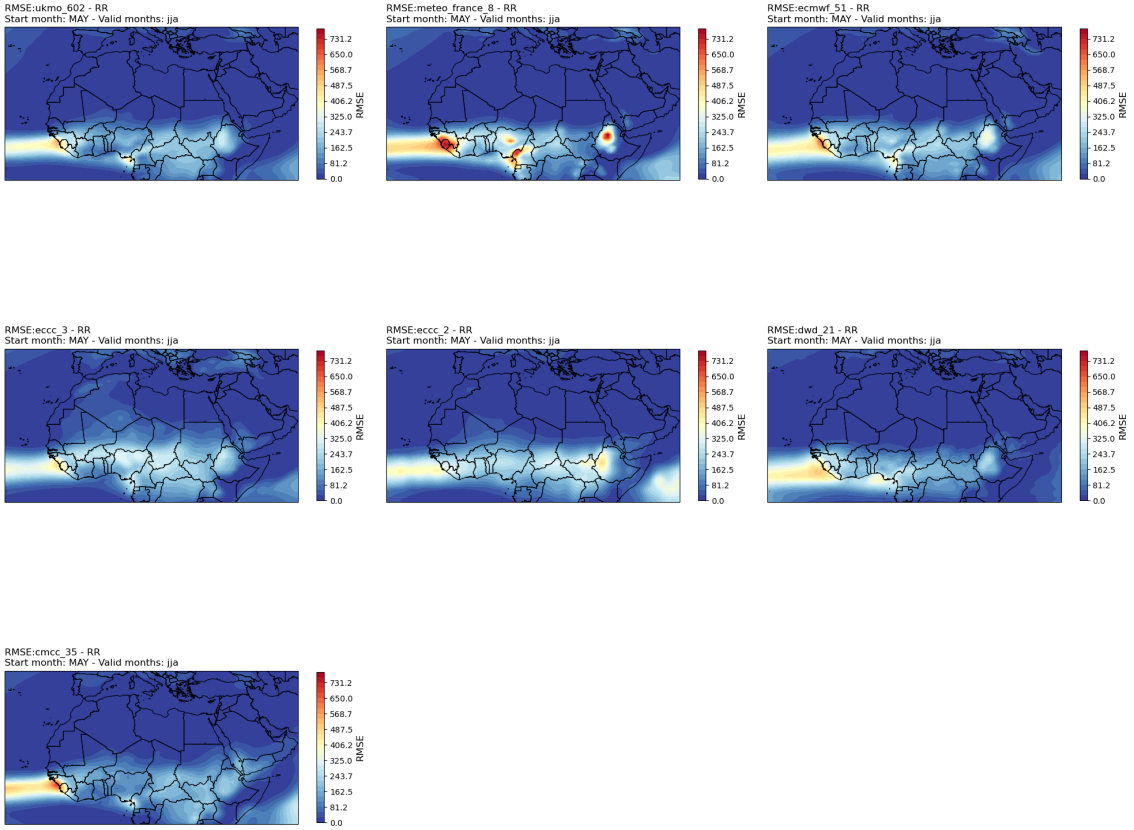

Figure 4.30: 3-months Rolling mean of RMSE in MENA Region for all centers JJA in mm

also for the spacial dimension, the RMSE stay stable and exhibit moderate performance for all centers. Thus, all models have almost the same skill and they are consistent with each other.

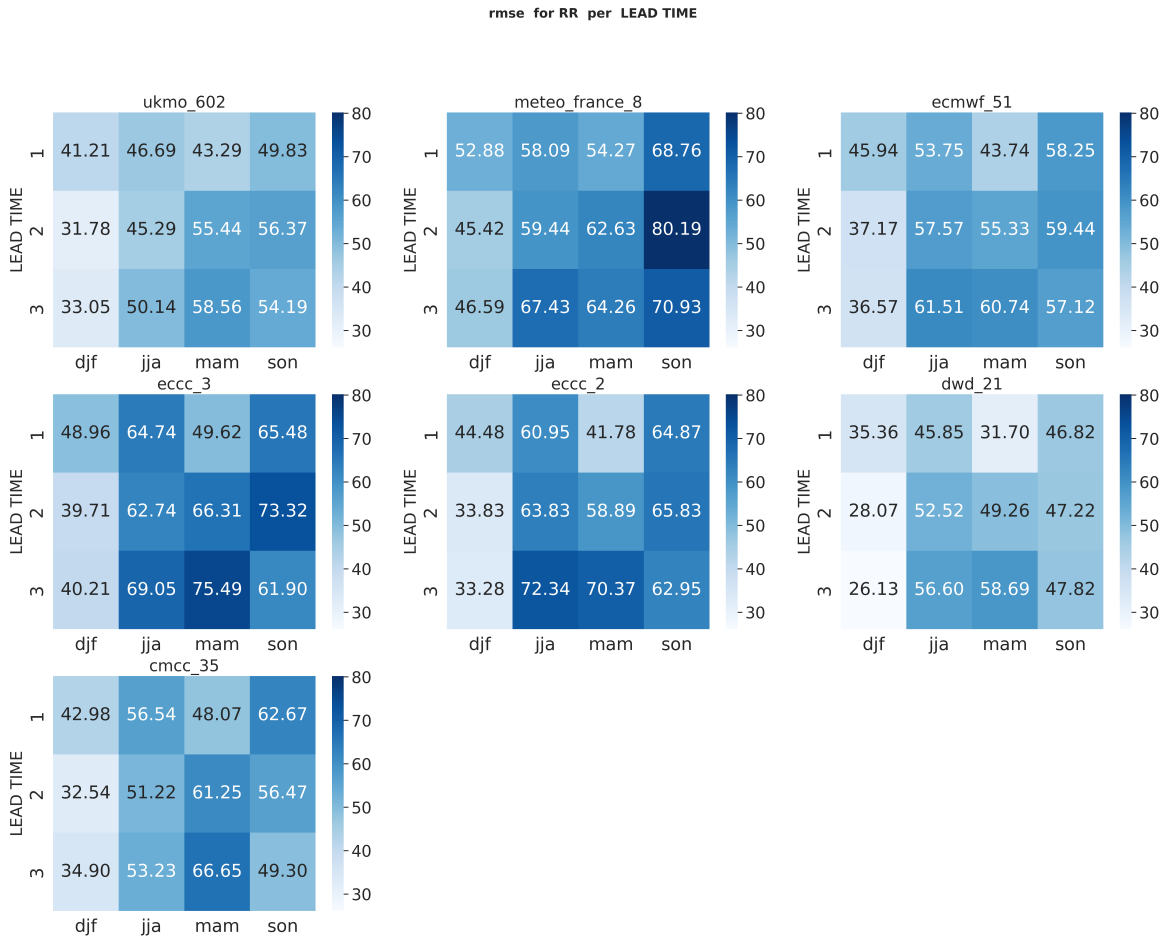

Figure 4.31: heatmap of RMSE For RR in mm

focus on North Africa :

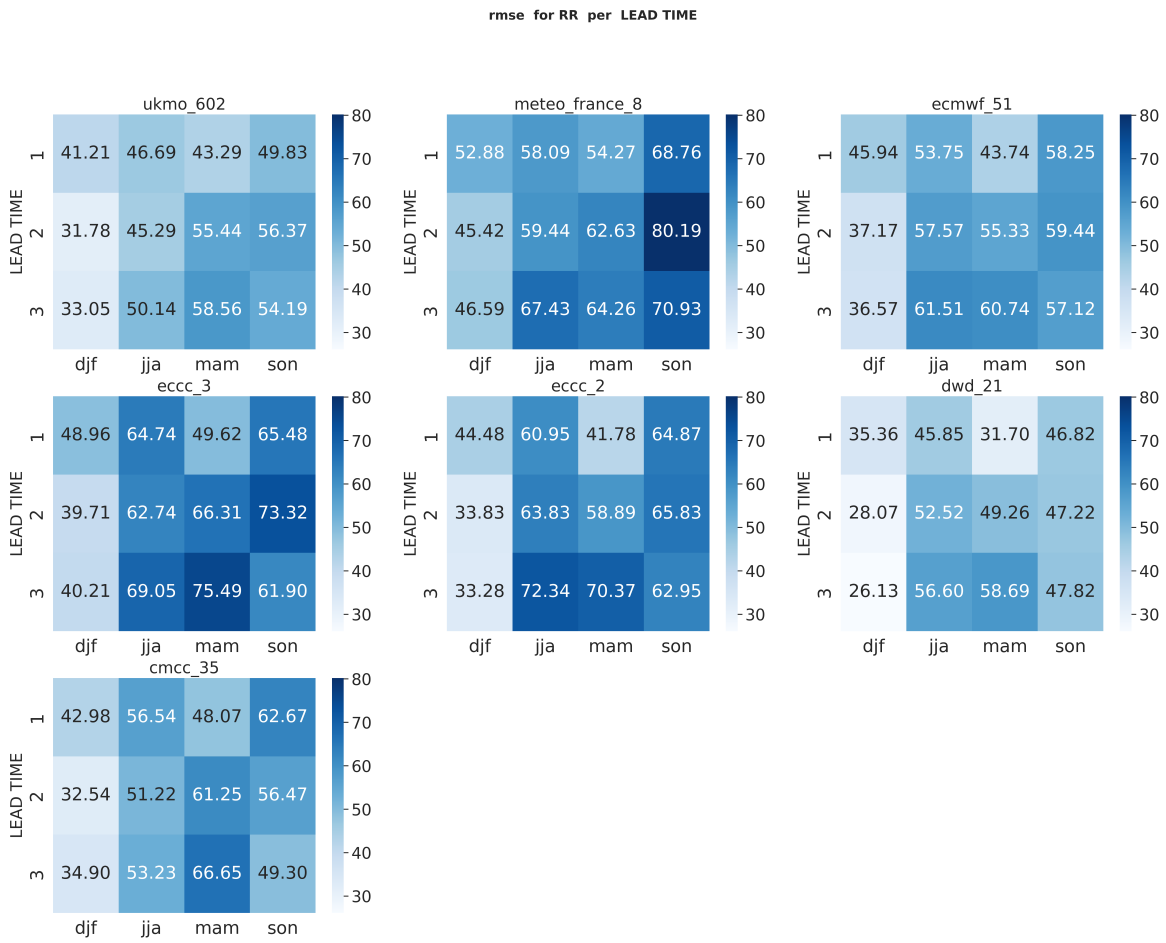

Figure 4.32: heatmap of RMSE For RR in mm (North Africa)

the RMSE is much better for North africa, the score is good over all lead-times and seasons. For centers, ***ecmwf, ukmo and dwd*** show very good performance.

Coefficient of Determination (R^2)

for precipitation, the R-SQUARED is very low, the maximum value is less than 0.1. However, the ecmwf is the best in term of R-SQUARED. for DJF,JJA and MAM the highest performance is in the first Lead-time, and it decrease along time, But for SON the best score is in the second Lead-time for all centers.

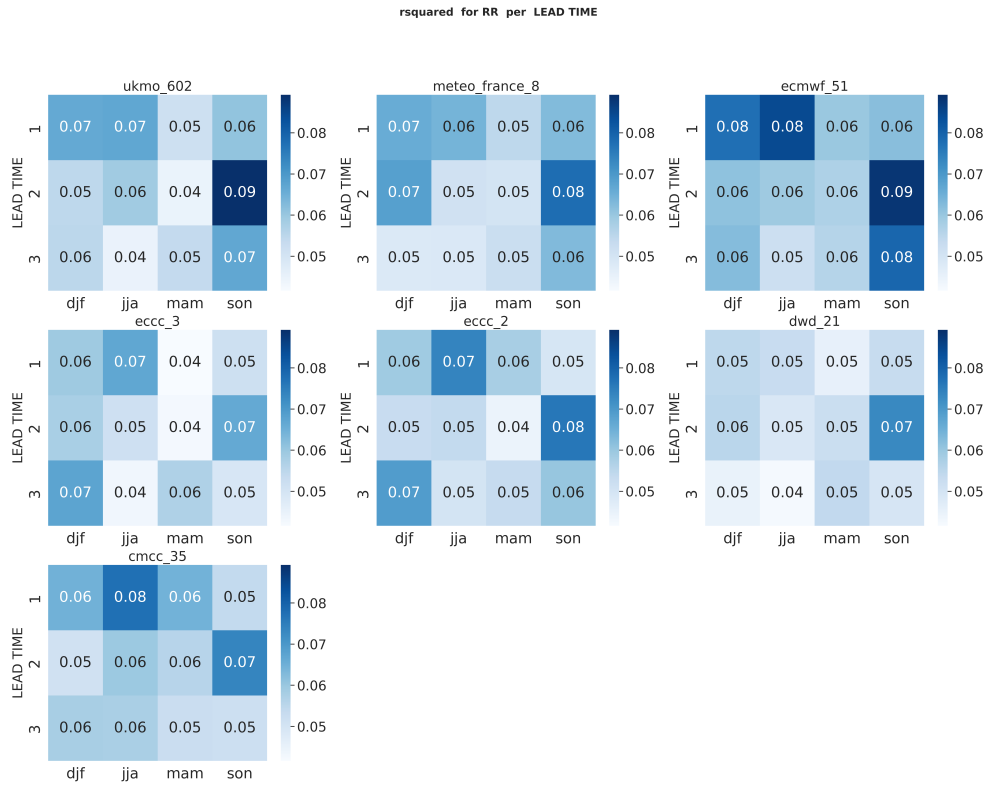

Figure 4.33: The Heatmap of rsquared for Precipitations in the mena region for every period ***(1 for perfect RSQUARED)***

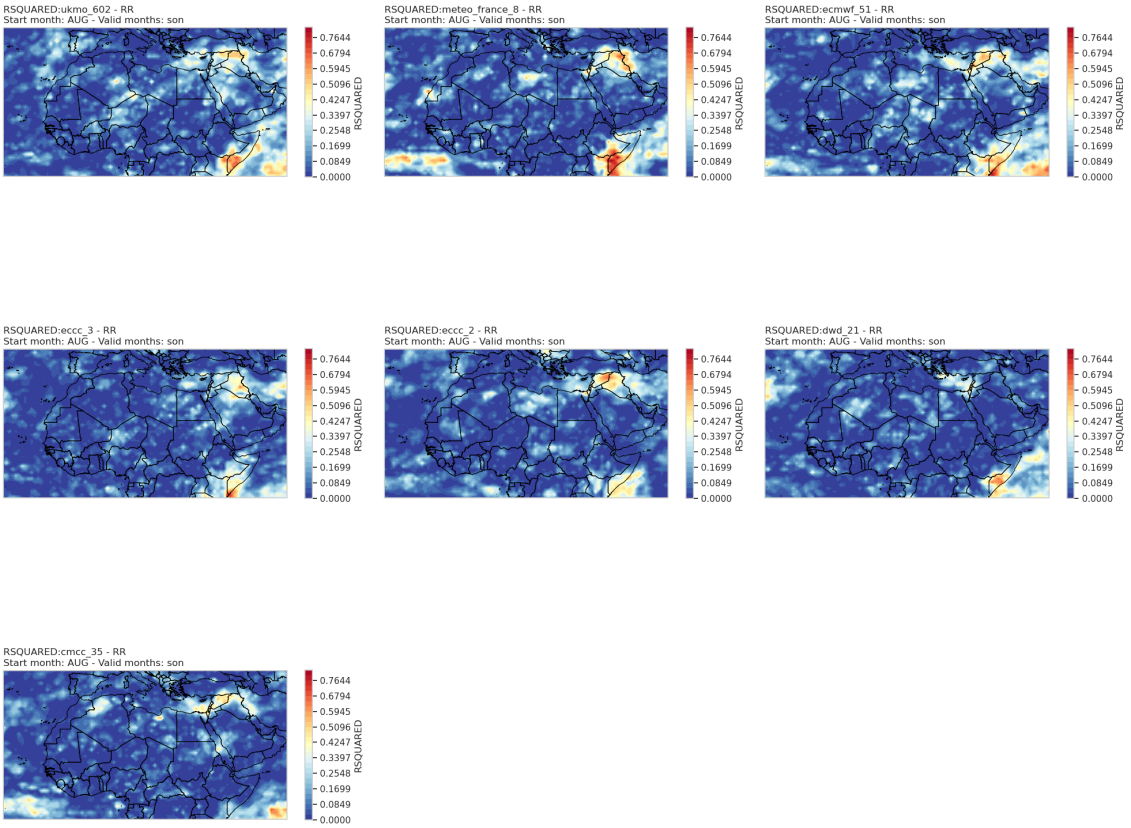

Figure 4.34: 3-months Rolling mean of RSQUARED in MENA Region for all centers SON

there is some isolated zones where the r-squared is good especially in Syria, Irak, Jordan ,Palestine and East Africa, this high performance is observed in all centers. For the rest of the MENA region the performance is very bad with score near to 0. Hence, there is no constant pattern for the R-SQUARED, the spacial variation is very high for all centers.

focus on North Africa
there is no big difference in North Africa.

4.2.2 Probabilistic Evaluation Metrics

The Brier Score (BS)

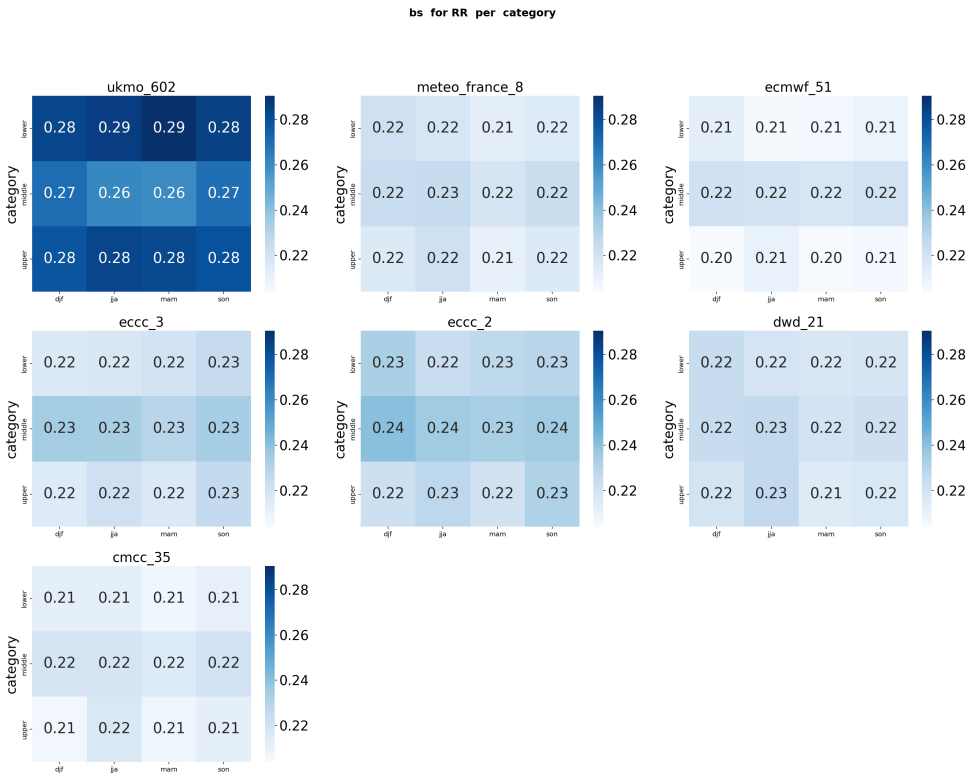

Figure 4.35: The Heatmap of Brier Score for each category . ***(0 represents perfect BS)***

for the analysis per category, we can see in the figure above that all centers exhibit good performance in term of Brier Score except the UMKO that shows moderate BS. Overall, the middle tercile shows lower performance (higher Brier Score) for all centers. the figure below shows the analysis per lead-time. the same result is found, but the ***ECMWF,METEO-FRANCE and CMCC-35*** are the best models in Brier Score for lead-time analysis.The performance stay stable along time which is a reliable signal. Despite the UKMO have the lower performance, it stays close to the other centers, the difference isn't so wide. In general, the performance stays stable over category, lead-time and space.

Figure 4.36: The Heatmap of Brier Score for lead-time. ***(0 represents perfect BS)***

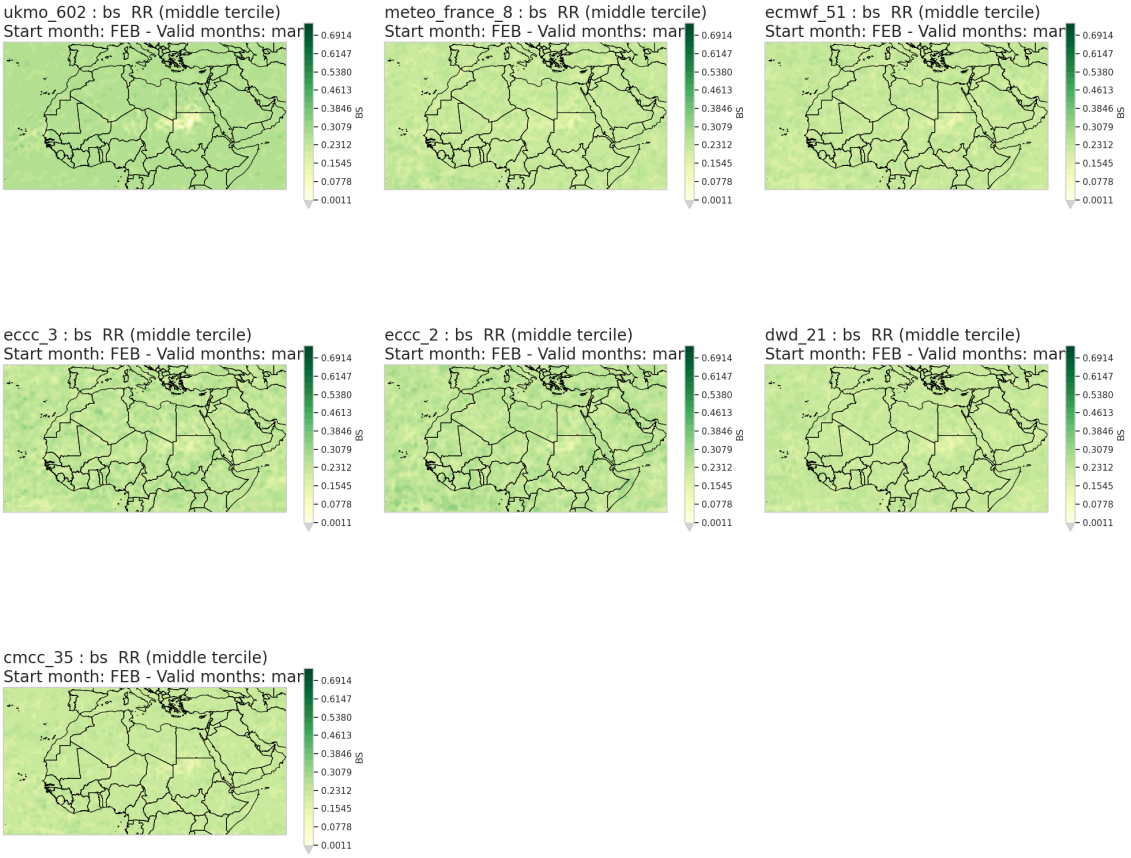

Figure 4.37: 3-months Rolling mean of Brier Score in MENA Region for all centers middle tercile MAM

the spacial distribution of the BS is homogeneous, the same performance across the MENA region, almost all centers perform well for all lead-times, for tercile there is a little lower performance for the middle tercile. Hint, for the other seasons the results are almost the same.

focus on North Africa:
there is no big difference in North Africa.

Reliability

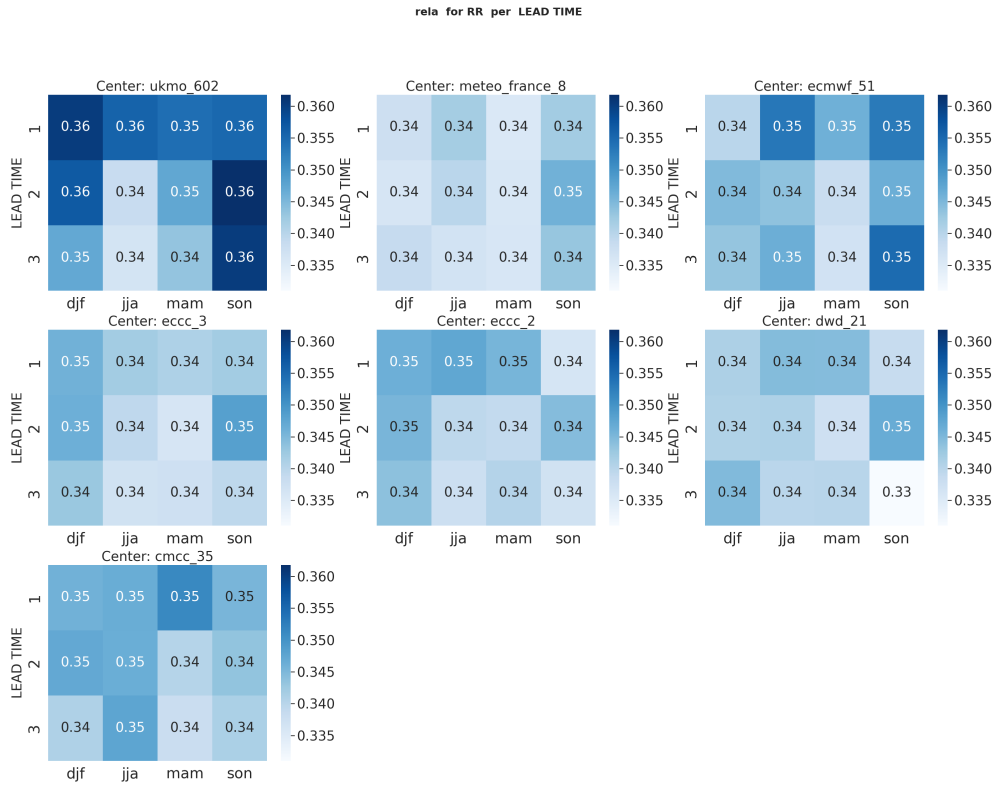

Figure 4.38: The Reliability Score . ***(0 means perfect Reliability)***

In the figure above, all centers demonstrate similar moderate performance in term of reliability. But deep analysis within the figure below, shows that UKMO has very bad performance, also we can distinguish three models that have the best reliability according to the reliability diagram, the centers are ***ECMWF,CMCC and METEO-FRANCE***. Hence,all centers give similar description of the reliability, also the stability along lead-time is a good indicator despite of the moderate results (0.3), we can rely on the models cited above because of the acceptable results and the stability along time.

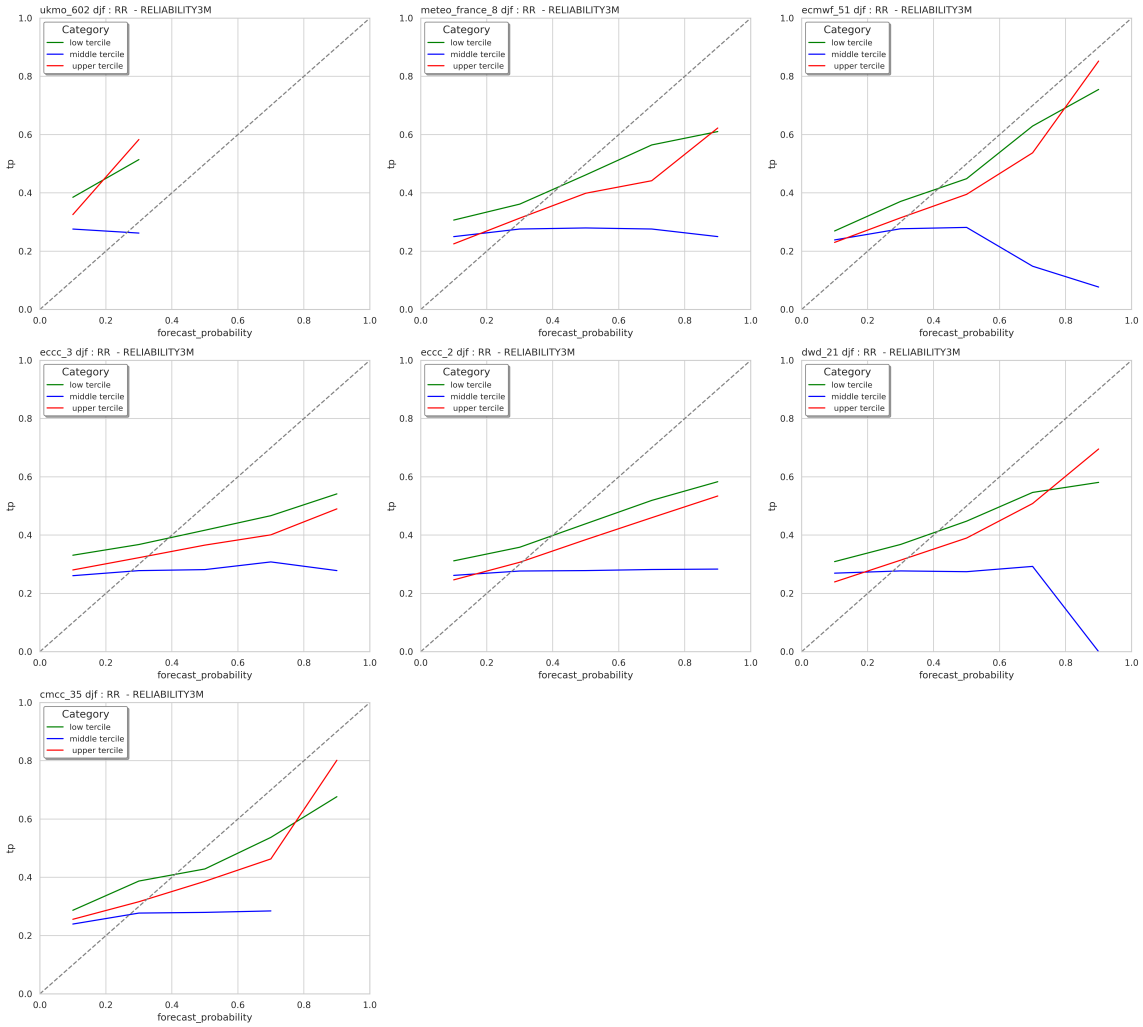

Figure 4.39: The 3-month rolling mean for Reliability DJF . ***Reliability is better in cases where the graphs are closer to the 45-degree line***

for the reliability diagram, all centers show moderate results, except for the ukmo that shows lower reliability.Thus, for the lower and upper terciles, models in general show good reliability, but for the middle tercile, this models aren't reliable. above all, ***ecmwf*** shows the highest performance for reliability.

focus on North Africa:
there is no big difference in North Africa.

The ranked probability score (RPS)

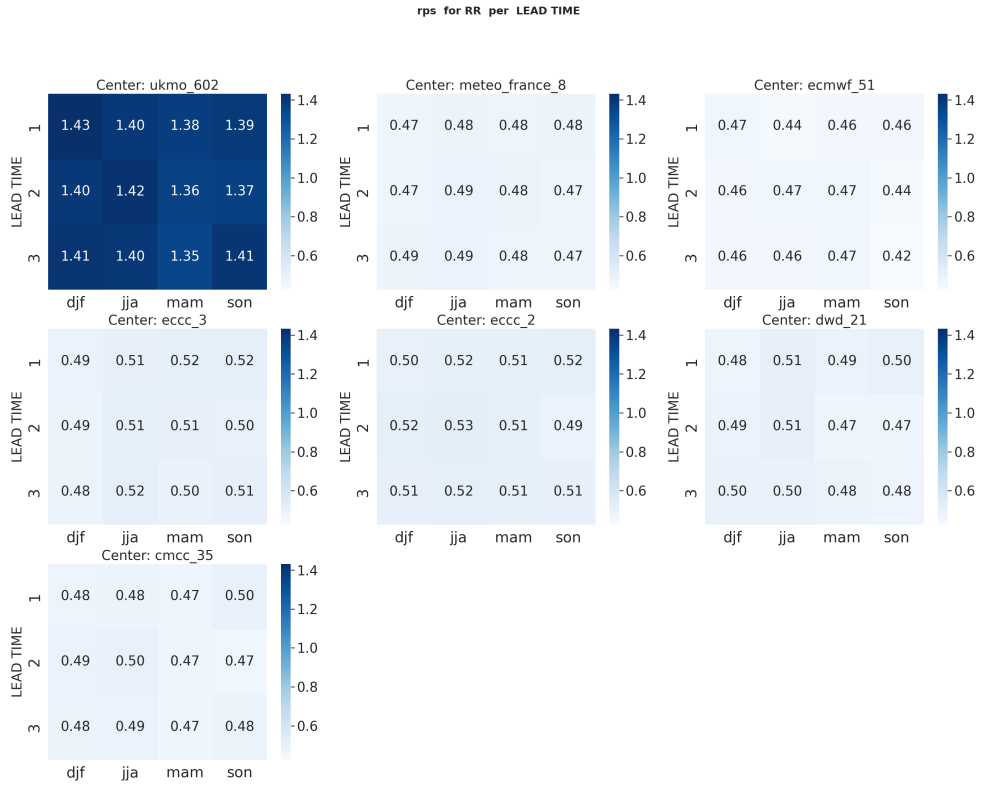

Figure 4.40: The Heatmap of RPS Score on MENA region for Precipitations . ***(0 means perfect RPS)***

In the figure above, all centers demonstrate moderate performance, except for UKMO, which shows noticeably lower performance.

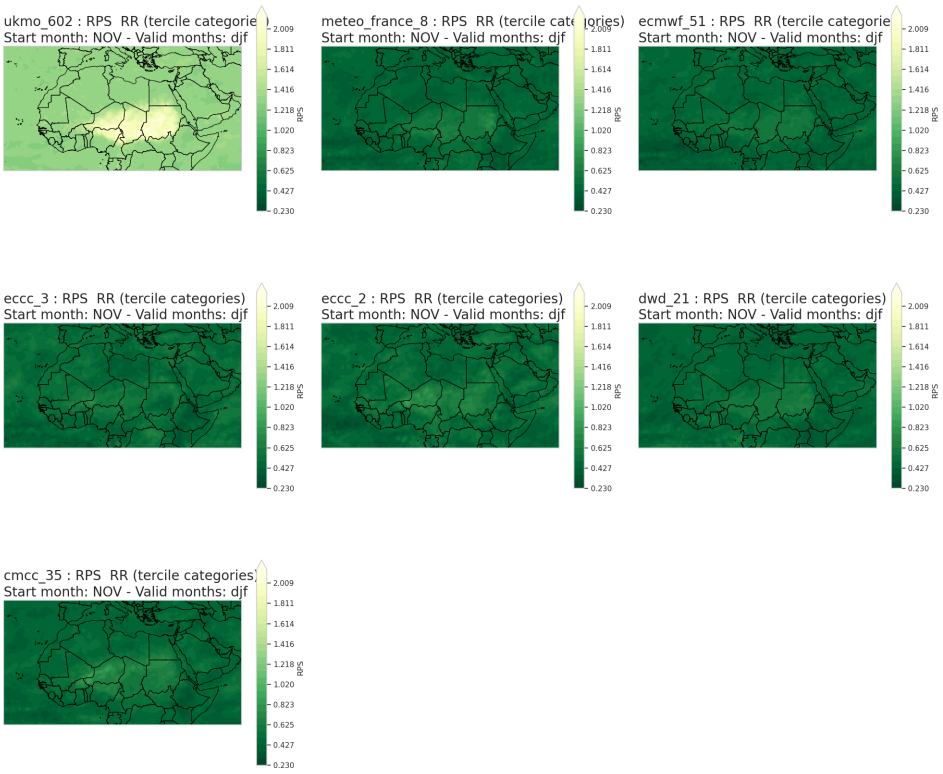

Figure 4.41: The RPS Score on MENA region for Precipitations DJF . ***(0 means perfect RPS)***

the spacial distribution of the RPS, is homogeneous, in all the region the score is good for all centers.Thus, the ukmo shows lower performance for this score.

focus on North Africa:
there is no big difference in North Africa.

Relative operating characteristics

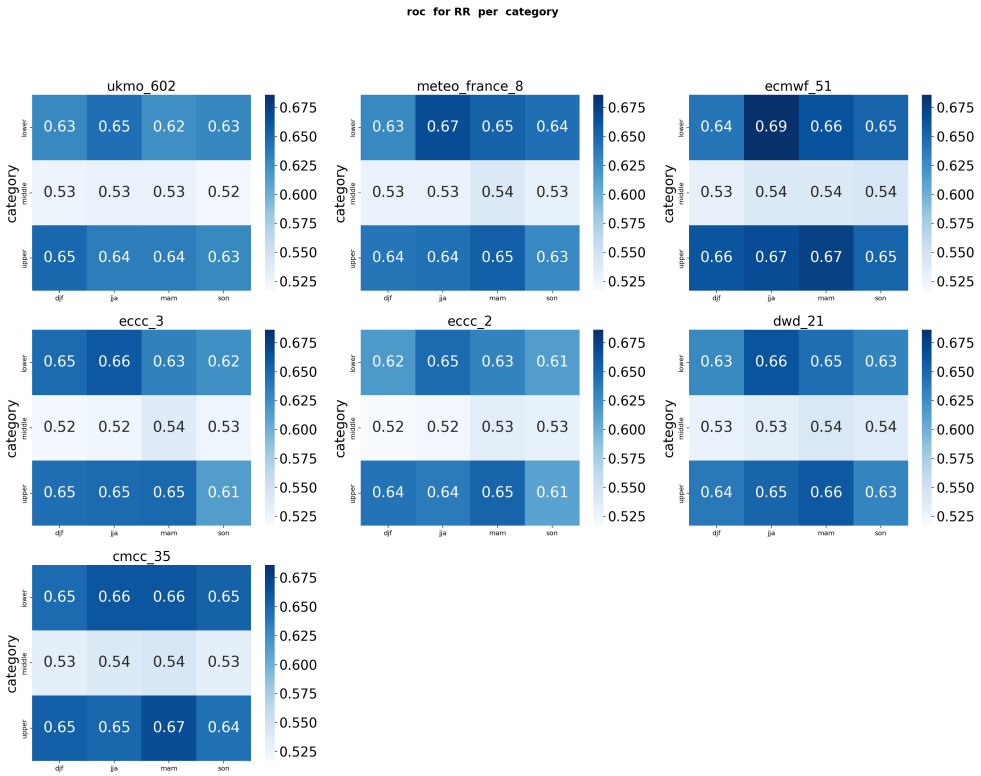

Figure 4.42: The Heatmap of ROC Score for each category . ***(1 means perfect ROC)***

In the figure above, it is evident that all centers exhibit similar performance levels. However, the middle tercile consistently achieves the lowest score.

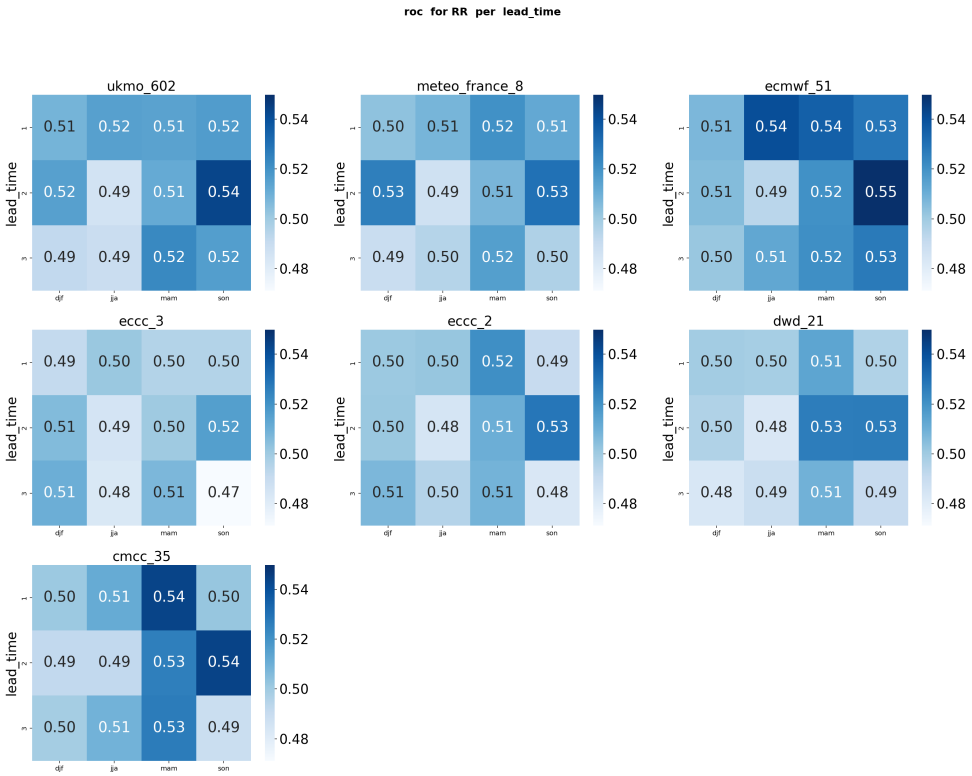

Figure 4.43: The Heatmap of ROC Score for lead-times. ***(1 means perfect ROC)***

for the ROC score, all centers show similar good performance, in general the best score is observed for tha first lead-time, except for the SON season where the best performance is for the second lead-time.

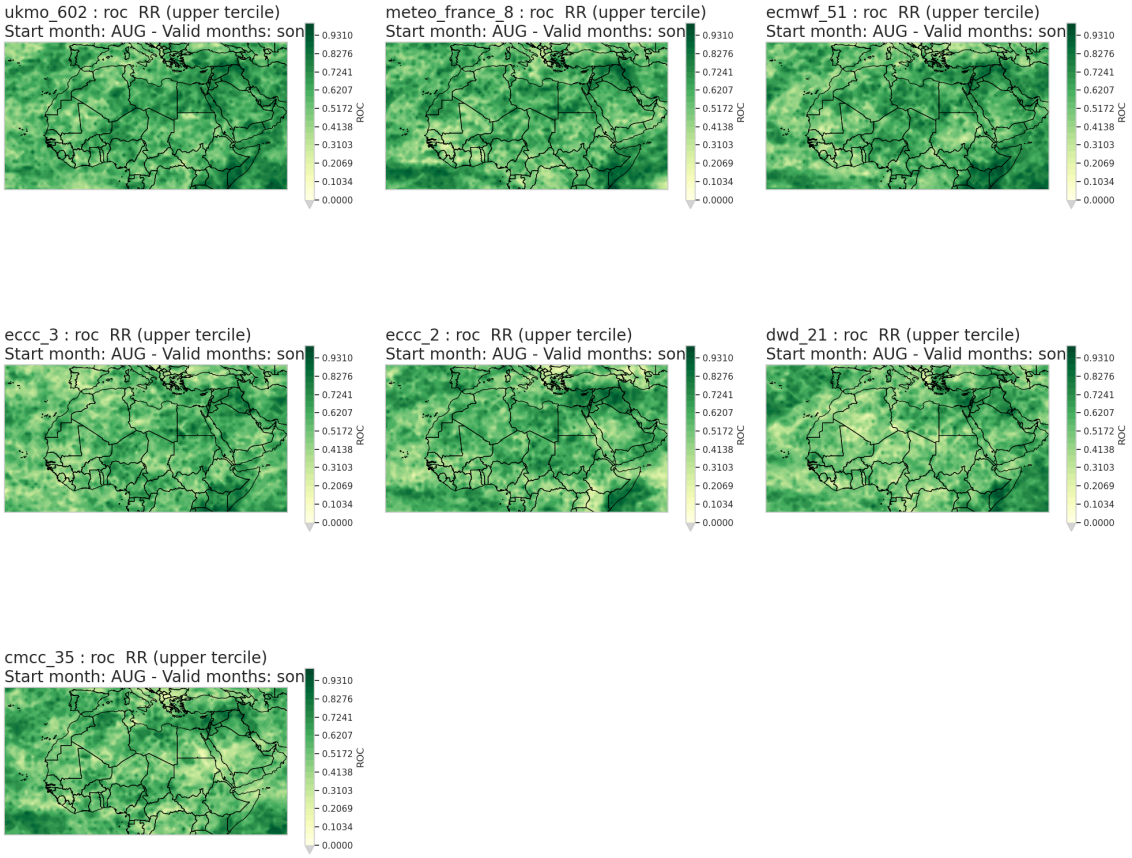

Figure 4.44: The ROC Score Upper tercile SON . ***(1 means perfect ROC)***

the spacial distribution of the roc score confirms an important result. For precipitation all centers shows better performance for the East of Africa, Irak,Syria,Jordan and Palestine. The performance in this zone is very high (score near to 1). For the rest of MENA region the performance is similar, with moderate to good score.

focus on North Africa:
there is no big difference in North Africa.

Relative operating characteristics Skill Score

In the figure above, the ECMWF exhibit the best performance for all terciles and periods. However, we should notice that the performance is very low for the middle tercile in all centers. For the analysis along time, the performance is so low, the best performance is in the first lead-time, except the SON which shows the best performance for the second lead-time. Above all, the ***ecmwf***

shows the best performance.

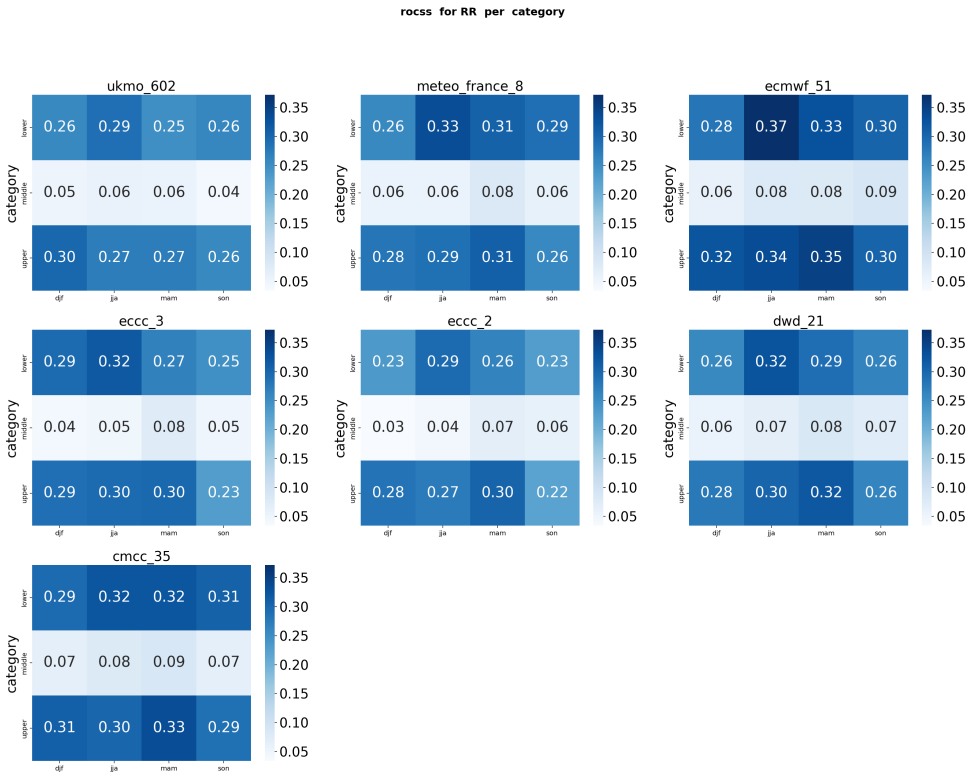

Figure 4.45: The ROCSS Score for each category . ***(1 means perfect ROCSS)***

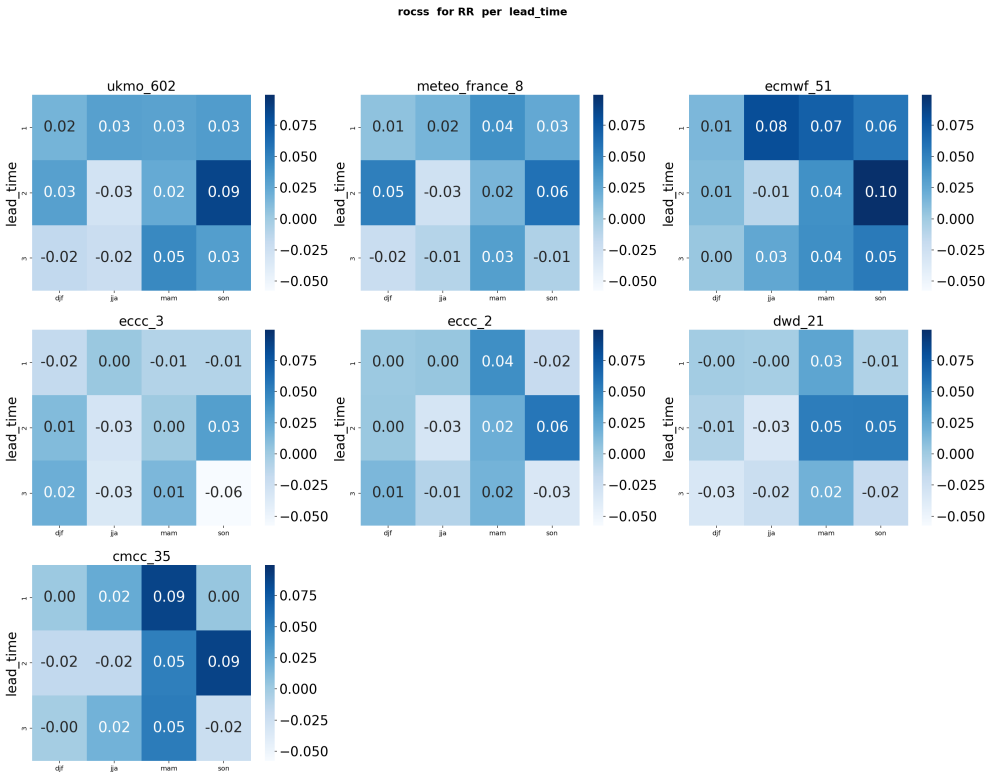

Figure 4.46: The average of ROCSS Score on all categories . ***(1 means perfect ROCSS)***

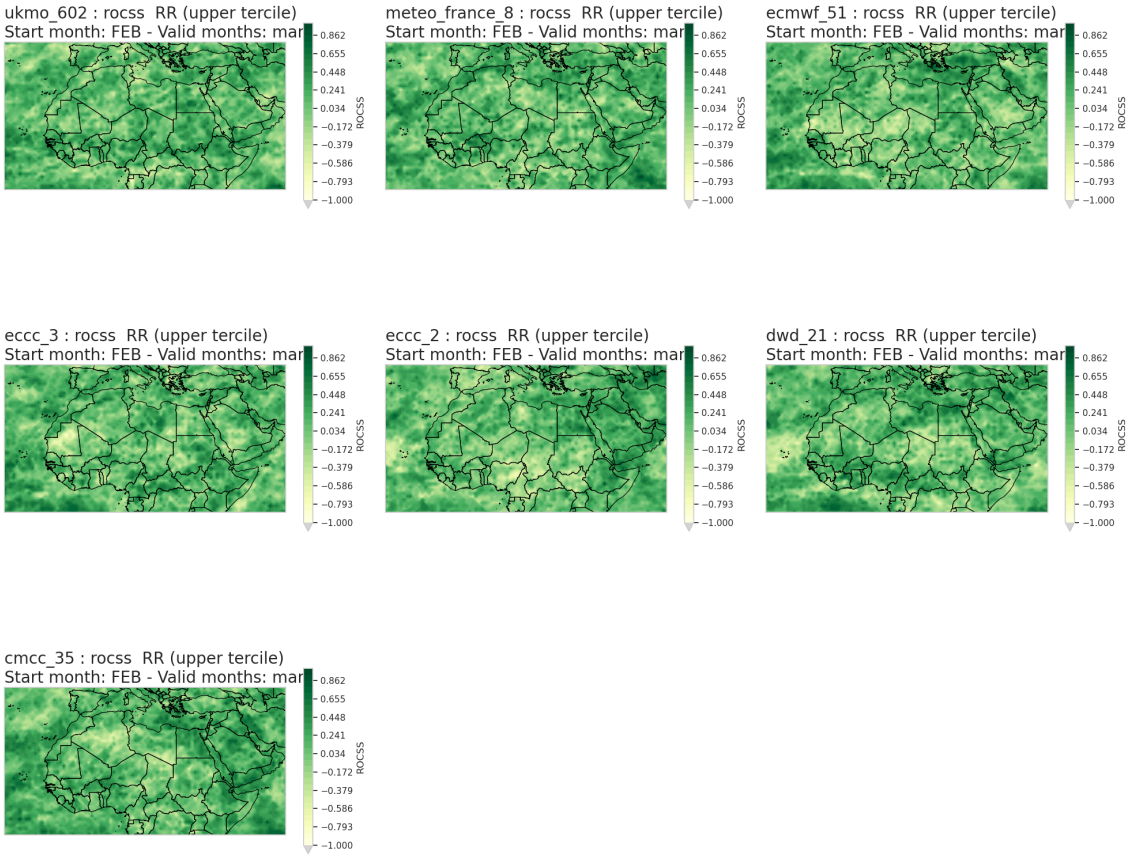

Figure 4.47: The ROC Skill Score Upper tercile MAM . ***(1 means perfect ROCSS)***

the spacial distribution of the ROCSS, shows that all centers are consistent for this score. The spacial distribution isn't clear, there is a high spacial variability.

focus on North Africa:

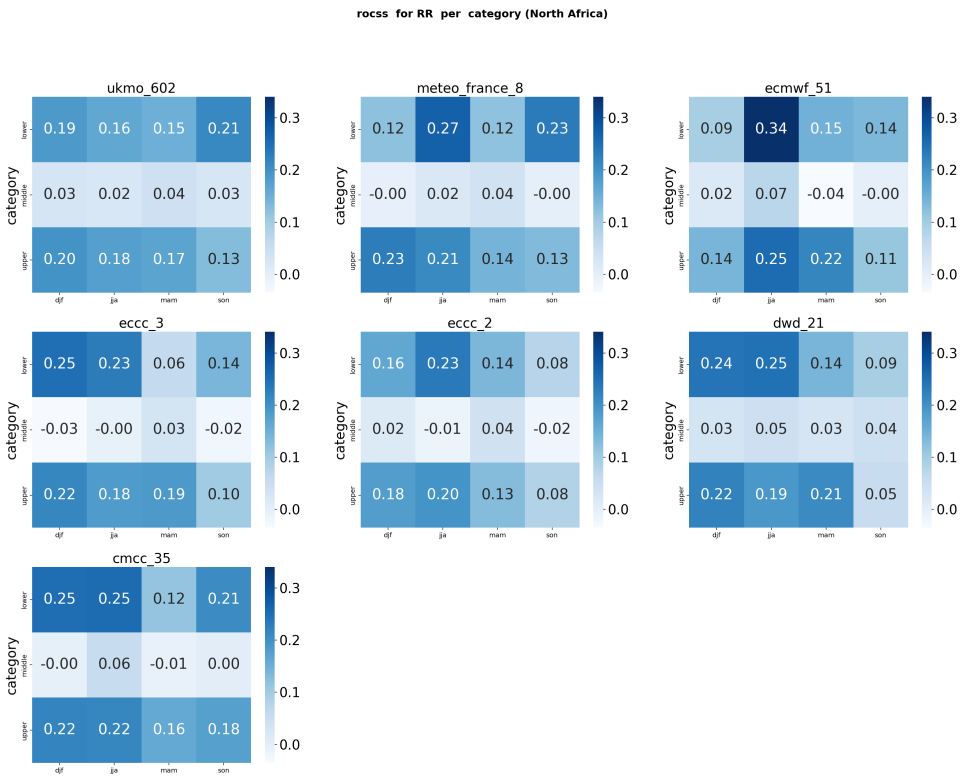

Figure 4.48: The ROCSS Score for each category North Africa . ***(1 means perfect ROCSS)***

Figure 4.49: The average of ROCSS Score on all categories . ***(1 means perfect ROCSS)***

the rocss for North Africa is in general lower, thus the performance is less accurate.

summary

Metric	Focus	What it Measures	Dependent on Observed Outcomes?	Visualization/Tools
Reliability	Probabilities match observed frequencies	Calibration of probabilities	Yes	Reliability diagram
Discrimination	Differentiating between outcomes	Ability to distinguish events from non-events	Yes	ROC curve, AUC
Sharpness	Boldness of probabilities (away from average)	Confidence of the forecast	No	Histogram of forecast probabilities
Resolution	Informativeness and variability of forecast	Ability to provide specific, useful info	Yes	Brier Score decomposition

Table 4.1: Key differences between reliability, discrimination, sharpness, and resolution in seasonal forecasting.

List of Figures

List of Tables